Journal of Geophysical Research

VOLUME 64 FEBRUARY 1959 NUMBER 2

THE SCIENTIFIC PUBLICATION
OF THE AMERICAN GEOPHYSICAL UNION

Journal of Geophysical Research

An International Scientific Publication

This Journal welcomes original scientific contributions on the physics of the earth and its environment.

Manuscripts should be transmitted to J. A. Peoples, Jr., Geology Department, University of Kansas, Lawrence, Kansas. Authors' institutions, if in the United States or Canada, are requested to pay a publication charge of $15 per page, which, if honored, entitles them to 100 free reprints.

Subscriptions to the *Journal of Geophysical Research* and *Transactions, AGU* are included in membership dues.

Non-member subscriptions, *Journal of Geophysical Research*.....$16 per calendar year, $2 per copy

Non-member subscriptions, *Transactions, AGU*.....$4 per calendar year, $1.25 per copy

Subscriptions, renewals, and orders for back numbers should be addressed to American Geophysical Union, 1515 Massachusetts Ave., Northwest, Washington 5, D. C. Suggestions to authors are available on request.

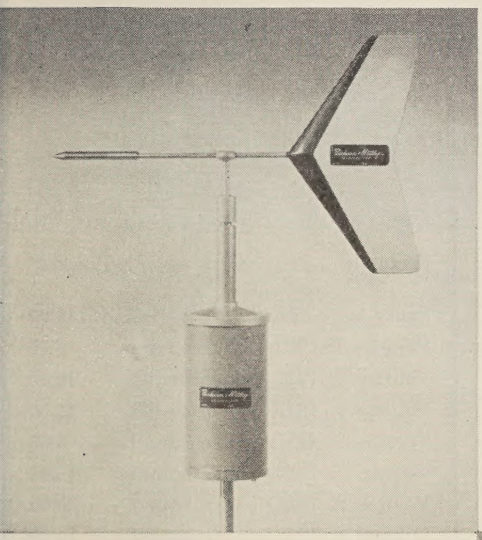
Advertising Representative: Howland and Howland, Inc., 114 East 32nd St., New York 16, N. Y.

Beginning with the January 1959 issue (Vol. 64, No. 1) the *Journal of Geophysical Research* is published monthly by the American Geophysical Union, 1515 Massachusetts Ave., Northwest, Washington 5, D. C., with the support of the Carnegie Institution of Washington and the National Science Foundation. The new monthly combines the type of scientific material formerly published in the bi-monthly *Transactions, American Geophysical Union*, and the quarterly *Journal of Geophysical Research*. The *Transactions, American Geophysical Union* will continue as a quarterly publication for Union business and items of interest to members of the Union.

Second-class postage paid at Richmond, Virginia.

OR SYSTEM
PPLICATION

Wind Transmitters



Highly adaptable, Beckman & Whitley Climate Survey Wind-Speed and -Direction Transmitters have wide applicability. They serve not only as elements in Beckman & Whitley Wind-Speed and -Direction Recorders, but also as basic standardized units for scientific weather measuring systems of special design and scope—involving telemetering, tape recording, other data-handling techniques.

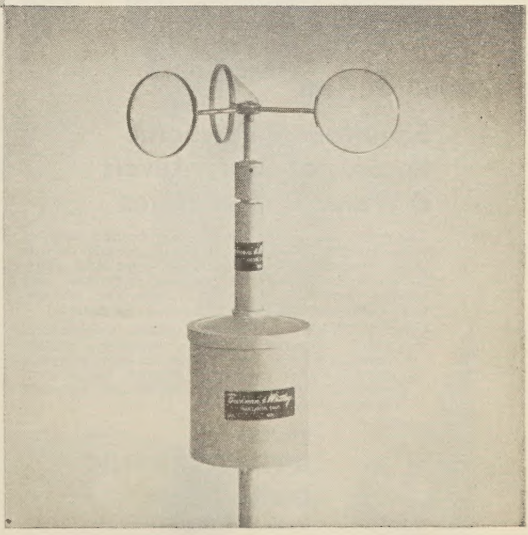
←

WIND COMPONENT TRANSMITTERS with low threshold, providing sine and cosine vector resolution for wind-component determination. Also standard linear wind-direction transmitters in the same basic design.

→

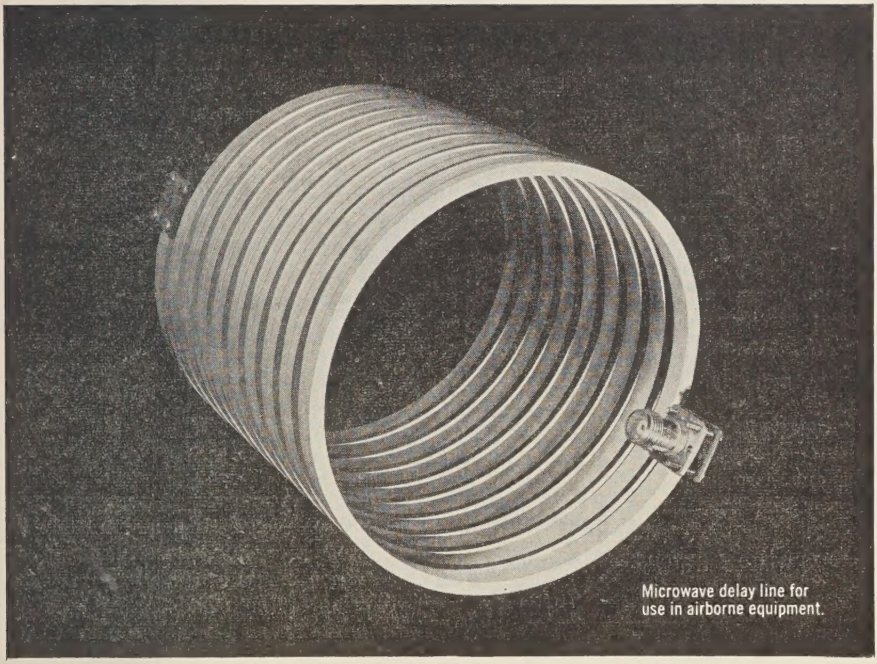
WIND SPEED TRANSMITTERS based on dragfree light-beam chopper design, available in standard types providing one, two, four, and 100 pulses per revolution. Complete series housed in identical environment-proved package with triple-labyrinth dust-seals, selected and specially-processed low friction bearings. Extremely rapid transient response —guaranteed threshold three-quarters mile per hour.

Send for details on these standard units, or, if you have special problems, for recommendations on other instruments or special adaptations.



Beckman & Whitley

instruments for scientific meteorology



Microwave delay line for use in airborne equipment.

MICROWAVE RESEARCH

The expanding role of electronic equipment in modern military operations has given high priority to microwave research. No field today offers greater challenge to the scientist and engineer.

In support of current electronic countermeasures programs and in anticipation of future systems requirements, Ramo-Wooldridge Division is engaged in microwave research to develop new techniques and to refine conventional components.

Research is under way at Ramo-Wooldridge for new methods and new designs to reduce substantially the over-all size, weight and complexity of electronic equipment for both airborne and ground-based uses.

For example, the low-loss delay line in the photograph above was designed, developed and manufactured by Ramo-Wooldridge for use in airborne equipment. Packaged for use in the system for which it was designed, this miniature ceramic unit weighs less than two pounds. It replaces a component which weighed more than twenty pounds and occupied more than five times as much volume.

Special opportunities exist for those with qualified experience in microwave research—in technique evaluation, component development, and design of such systems equipment—at Ramo-Wooldridge.

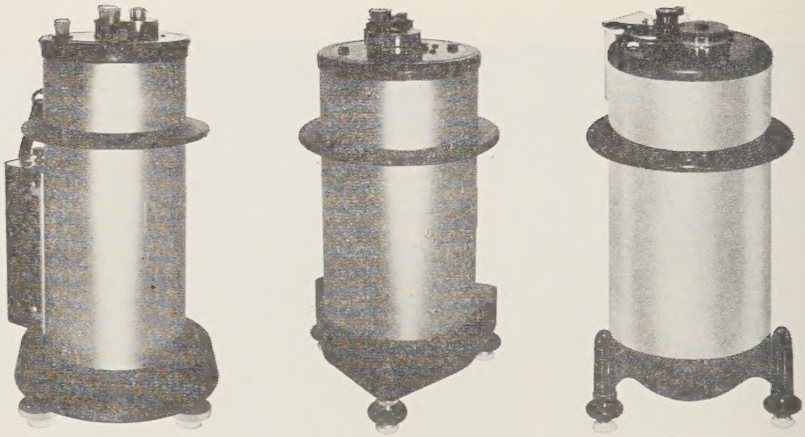
Engineers and scientists are invited to explore openings at Ramo-Wooldridge in:

- Electronic Reconnaissance and Countermeasure Systems
- Infrared Systems
- Analog and Digital Computers
- Air Navigation and Traffic Control
- Antisubmarine Warfare
- Electronic Language Translation
- Information Processing Systems
- Advanced Radio and Wireline Communications
- Missile Electronics Systems

RAMO-WOOLDRIDGE

P. O. BOX 90534, AIRPORT STATION • LOS ANGELES 45, CALIFORNIA

a division of ***Thompson Ramo Wooldridge Inc.***



the Master *the Prospector* *the Educator*

Now, you have a choice of three world-famous

TI WORDEN GRAVITY METERS

The MASTER retains the true portability and reliability for which the Worden Meter set a world precedent. However, the MASTER goes even further, having more demanding specifications plus a unique low-power temperature-stabilizing feature which together upgrade both the quantity and quality of gravity data. With the MASTER, in either severe or moderate temperature conditions, the absolute minimum number of base ties are required due to the positive linear drift, giving you a maximum production rate of the most accurate data at a reduced operating cost. The MASTER is truly the finest gravity meter available . . . with or without the temperature stabilizer in operation. Both the MASTER and PROSPECTOR have a gearless top reading dial which gives greater operator convenience and minimizes human error.

The PROSPECTOR has set reliability standards for gravity meters the world over. During the past ten years, over 450 of these gravity meters have been placed in use . . . this number exceeds the total of all other types combined. As with TI Worden Meters now in use, the PROSPECTOR will continue to provide exacting results in normal gravity programs. This is assured and even enhanced in that tighter manufacturing specifications have resulted in better temperature compensation and improved accuracy.

The EDUCATOR is designed to meet the needs of educational institutions, company training programs and surveys allowing for wider tolerances. This meter contains many of the outstanding features that have established Worden superiority, yet fulfills an increasing need for a reliable meter in limited budget projects.

Write today for complete information . . .
specify Bulletin G-205.

TEXAS INSTRUMENTS
INCORPORATED
INDUSTRIAL INSTRUMENTATION DIVISION
3609 BUFFALO SPEEDWAY • HOUSTON, TEXAS • CABLE: HOULAB

OTHER TI/IID PRODUCTS

- Seismic Systems
- DATA-GAGE Measurement and Control Systems
- "recti/riter" Recorders and Accessories
- Automatic Test Equipment

203

Journal of
GEOPHYSICAL RESEARCH

OLUME 64 FEBRUARY, 1959 No. 2

Arctic Measurements of Electron Collision Frequencies in the D-Region of the Ionosphere

J. A. KANE

U. S. Naval Research Laboratory, Washington, D. C.

Abstract—Electron collision frequencies in the *D*-region were measured with the aid of rockets launched during polar blackouts at Fort Churchill, Canada (58° N.). During the blackouts the ionization content of the *D*-region was large enough to cause a strong altitude-dependent absorption of a 7.75 Mc/s CW signal, radiated from the rocket. The refractive indices and the difference absorption of the two magneto-ionic components of this signal were simultaneously measured versus altitude. The electron collision frequency profile was then determined from these measurements.

The results from two midday rocket flights, one made in July, 1957 and the other in November, 1956, indicate that the electron collision frequencies in the Arctic *D*-region are lower by a factor of three than the standard values calculated by *Nicolet* [1957]. At 70 km the measured value of the collision frequency is $(1.0 \pm 0.3) \times 10^7$ sec^{-1}.

I. INTRODUCTION

Many experiments have been performed from ıe earth's surface to determine electron collision equencies for the lower ionosphere. The xperiments which are based upon ionosphere eflection properties or upon ionospheric cross ıodulation effects are necessarily indirect and in eneral quite complex. The published results ıow a wide variation. The situation has been eviewed by *Nicolet* [1953], who more recently 957] has calculated, on the basis of the nitrogen oss section measurements of *Anderson* and *oldstein* [1956], a table of theoretical values of ectron collision frequency for the 50 to 120 km titude interval.

By transmitting CW radio signals from a ıcket, it is possible to measure in a relatively mple manner the electron collision frequency, ovided certain ionospheric conditions exist ıring the time of the rocket flight. Such conıtions were obtained during the flight of the RL Aerobee-Hi rocket NN3.08F, fired at Fort hurchill, Canada, on July 4, 1957. (The ectron density measurements obtained from this flight are reported in a separate paper by J. C. Seddon and J. E. Jackson.) The firing was made at 12h 16m CST during a polar blackout in which the ionization content of the *D*-region was large enough to cause a strong altitude-dependent attenuation of the 7.75 Mc/s radio signal used in the experiment. This measurement, combined with a simultaneous measurement of the altitude dependence of the electron density, determines the electron collisional frequency versus altitude.

II. THEORY AND METHOD

The Appleton-Hartree expression for the refractive index for the case of propagation along a direction essentially parallel to the earth's magnetic field reduces to

$$\left(n - j\frac{c}{\omega}\kappa\right)^2 = 1 - \frac{X}{(1 \pm Y) - jZ} \quad (1)$$

where

n = real part of the refractive index
κ = index of absorption
c = velocity of light in vacuo
ω = exploring wave angular frequency

$X = 81N_e/(\omega/2\pi)^2$
N_e = electron density
ωY = electron angular gyro frequency corresponding to earth's magnetic field component in direction of propagation
$\omega Z = \nu$
ν = collision frequency of electrons with neutral particles
$(1 + Y)$ is associated with the ordinary propagation mode
$(1 - Y)$ is associated with the extraordinary propagation mode

Equating the real and imaginary parts of (1) yields the following equations to the first order in X

$$n = 1 - \frac{X}{2} \frac{(1 \pm Y)}{(1 \pm Y)^2 + Z^2} \tag{2}$$

$$2n\left(\frac{c\kappa}{\omega}\right) = \frac{XZ}{(1 \pm Y)^2 + Z^2} \tag{3}$$

from which, for n close to unity:

$$\frac{c\kappa}{\omega} = \frac{Z}{(1 \pm Y)}(1 - n) \tag{4}$$

Designating the ordinary and extraordinary propagation modes by the subscripts 0 and x respectively, the difference absorption $\kappa_x - \kappa_0$ becomes

$$\kappa_x - \kappa_0 = \frac{\omega}{c} Z\left[\frac{1 - n_x}{1 - Y} - \frac{1 - n_0}{1 + Y}\right] \tag{5}$$

Also, for later use a function $q(Z, Y)$ is defined as

$$q(Z, Y) \equiv \frac{1 - n_x}{1 - n_0} = \frac{1 - Y}{1 + Y}\left[\frac{(1 + Y)^2 + Z^2}{(1 - Y)^2 + Z^2}\right] \tag{6}$$

In the standard NRL propagation experiment [*Seddon*, 1953] two CW signals are transmitted from the rocket. One signal is 7.75 Mc/s, the other is the sixth harmonic of this frequency. On the ground the received low frequency signal is frequency multiplied by 6 and compared with the ordinary component of the received high frequency signal. This process is performed separately for both polarization components of the 7.75 Mc/s signal. The resulting beat frequency is given for the case of a non-rolling rocket by

$$F_B = 6f\frac{v}{c}(1 - n) \tag{7}$$

where

f = 7.75 Mc/s
v = rocket velocity (determined independently of the experiment)
n = refractive index of the medium at the rocket position.

In practice a small rocket roll rate is present which is in general less than 1 cps. The resulting variations in the antenna orientation alter both the received high and low frequency signals by cps. However, since the received low frequency signal is frequency multiplied by 6 before comparison, the resultant beat note for the ordinary mode is

$$F_{B_0} = 6f\frac{v}{c}(1 - n_0) + 5\rho \tag{8}$$

while the beat note for the oppositely polarized mode becomes

$$F_{B_x} = 6f\frac{v}{c}(1 - n_x) - 7\rho \tag{9}$$

Averaged over one roll period, the value of ρ is the same as the average value of the rocket roll, the instantaneous value of ρ, however, varies in a manner determined by the relative geometry of the transmitting and receiving antennas [*Jackson*, 1954]. The instantaneous value of ρ is not measured. However, by means of (8) the rocket roll term can be eliminated from (8) and (9) to give

$$\frac{F_{B_x} + 1.4F_{B_0}}{q + 1.4} = \frac{6fv}{c}(1 - n_0) \tag{10}$$

During the experiment, in addition to the beat notes F_{B0} and F_{Bx}, a continuous recording was also made of the signal strengths of the two polarization modes of the 7.75 Mc/s signal. As a function of rocket altitude z (not to be confused with $Z \equiv \nu/\omega$), the signal strength $E_{0,x}$ received on the ground can be written for either polarization mode as

$$E_{0,x} = \frac{E_{0,x}'}{z} \exp\left(-\int_0^z \kappa_{0,x}\, dz\right) \tag{11}$$

The difference absorption $\kappa_x - \kappa_0$ is then

$$\kappa_x - \kappa_0 = \frac{d}{dz} \ln \frac{E_0}{E_x} \tag{12}$$

where the term

$$\frac{d}{dz} \ln \frac{E_0'}{E_x'}$$

has been assumed to be negligible. This point is considered in Section IV. Substitution of (12) and (10) in (5) yields, with the aid of (6),

$$\frac{6v \dfrac{d}{dz} \ln \left(\dfrac{E_0}{E_x}\right)}{2\pi(F_{B_x} + 1.4F_{B_0})} \tag{13}$$

$$= \frac{Z}{1 - Y} \left\{ \frac{q(Z, Y) - \left(\dfrac{1 - Y}{1 + Y}\right)}{q(Z, Y) + 1.4} \right\} \equiv Q(Z, Y)$$

With the definition of $q(Z, Y)$ given by (6), it is possible to obtain a graph of the function $Q(Z, Y)$, shown in Figure 3. Over the altitude region involved in the experiment, the magnetic field term Y can be considered a constant. Thus a measurement of the three experimental quantities, $\ln (E_x/E_0)$, F_{Bx}, and F_{B0} determines the value of $Q(Z, Y)$. From the graph of $Q(Z, Y)$ vs $Z = \nu/\omega$, two values of the electron collision frequency are determined, one of which is excluded by the consideration that ν must be a decreasing function of altitude.

III. Results

A. *Data*

During the rocket ascent through the altitude region involved, the flight record was free of the obfuscating effects of multiple path propagation or related phenomena. During this portion of the flight the rocket ascended along an essentially vertical straight line from the ground station. The experimental quantities F_{B0} and F_{Bx} are obtained from the flight record with accuracies of the order of a few per cent and require no further discussion. The quantity $F_{Bx} + 1.4F_{B0}$ is shown in Figure 1. From (8) and (9) it is seen that this combination of beat frequencies is independent of rocket roll and, apart from a rocket velocity term, dependent only on the refractive indices n_x and n_0.

Figure 2 is a plot of the experimental quantity $\ln (E_x/E_0)$ versus z from which a slope must be obtained. The spread on the experimental points represents the accuracy to which the flight record could be read. Through the experimental points is drawn a visually-fitted solid curve whose slope is obtained as an average over intervals of two kilometers.

From this slope and the values of the beat notes F_{B0} and F_{Bx} as read from the flight record,

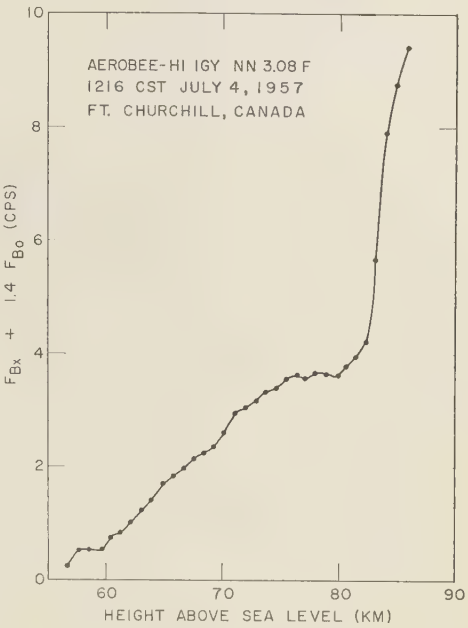


Fig. 1—The beat frequency combination $F_{Bx} + 1.4F_{B0}$ *vs* altitude.

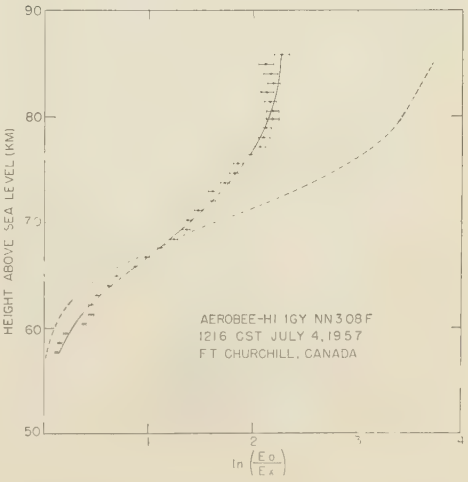


Fig. 2—The logarithmic ratio of the signal strengths of the two polarization modes *vs* altitude.

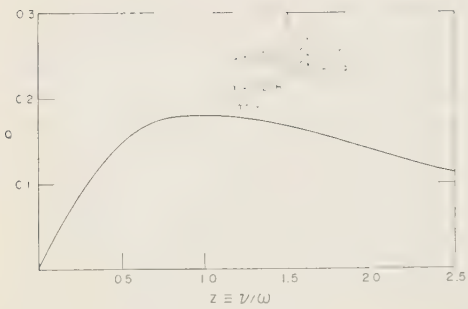


FIG. 3—The function $Q(Z, Y)$ *vs* $Z \equiv \nu/\omega$. $Y_{L,T}$ are the components of Y longitudinal and transverse to the direction of propagation.

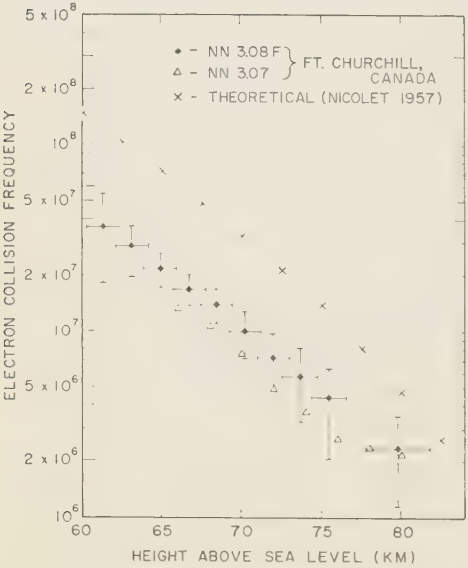


FIG. 4—Electron collision frequency *vs* altitude.

the collision frequencies are computed by means of (13) and the curve $Q(Z)$ of Figure 3. The resulting collision frequencies are listed in Table I and plotted in Figure 4. The horizontal spread on each point represents the altitude interval over which the collision frequency was averaged, while the vertical spread represents the uncertainty in the collision frequency arising from the uncertainty in the slope of the experimental points of Figure 2. The effect of uncertainties in the beat note frequencies is negligible by comparison. Also plotted in Figure 4 are the values of the collision frequencies

TABLE 1—*Electron collision frequency* ν *versus altitud* *z above sea level*

July 1957		November 1956*	
z (km)	$\nu(\text{sec}^{-1})$	z (km)	$\nu(\text{sec}^{-1})$
61.3	$(3.6 \pm 1.8) \times 10^7$		
63.1	$(2.8 \pm 0.8) \times 10^7$		
64.9	$(2.2 \pm 0.5) \times 10^7$		
66.7	$(1.7 \pm 0.3) \times 10^7$	66	1.3×10^7
68.4	$(1.4 \pm 0.3) \times 10^7$	68	1.1×10^7
70.2	$(1.0 \pm 0.3) \times 10^7$	70	7.5×10^6
72.0	$(7.2 \pm 2.6) \times 10^6$	72	4.9×10^6
73.7	$(5.7 \pm 2.5) \times 10^6$	74	3.7×10^6
75.5	$(4.3 \pm 2.1) \times 10^6$	76	2.6×10^6
		78	2.4×10^6
79.8	$(2.3 \pm 1.2) \times 10^6$	80	2.1×10^6

* The errors on the two flights are comparable.

as calculated by *Nicolet* [1957]. It is seen tha the altitude dependence is similar but tha there is disagreement by a factor of three whic is considerably larger than the indicated experi mental uncertainties. An examination of th sources of systematic errors failed to uncover an error of the magnitude required to explain disagreement of this size. Pfister [1954] ha calculated the corrections required in th Appleton-Hartree refractive-index formula whe the ratio of ν/ω approaches and exceeds unity In the altitude region involved in the presen experiment, however, the ratio ν/ω is smal enough so that these corrections are negligible

B. *Internal consistency*

To check the internal consistency of th results one can assume the collisional frequencie of Nicolet to be correct, which assumptio thereby determines the function $Q(Z)$, and b using the observed beat frequencies, comput the value of

$$\frac{d}{dZ} \ln \frac{E_2}{E_1}$$

from (13). The result, after integration, i plotted as the dashed line in Figure 2. Howeve when this computation is repeated, using for th collisional frequency a value equal to one third that of Nicolet, the result is in very good agree ment with the experimental points.

An additional check on the results was provide by a separate rocket flight (NN3.07) made a

Fort Churchill during a polar blackout on November 15, 1956 at 13h 32m CST. Although the data of that flight were of somewhat poorer quality, the results which were obtained during the descent of the rocket are in agreement with those of the present flight. These are listed in Table I and shown as triangles in Figure 4.

IV. Systematic errors

A. *Ion effects*

In the *D*-region, since the number of neutral particles far exceeds the number of ions, the frequency of electron-ion collisions is negligibly small compared with the frequency of electron-neutral particle collisions.

The effect of ions on the refractive indices can be carried through the algebra of equations (1) through (13) in a straightforward manner. The analysis shows, however, that the presence of ions would raise the experimental points of Figure 4 by a negligible amount for any ion density likely to be encountered in the *D*-region. The unlikely ion density of 10^7 ions/cc would increase the experimental collision frequencies by not more than 10 pct.

B. *Rocket induced disturbances*

A source of systematic error inherent in any rocket experiment is the fact that the rocket disturbs the medium through which it travels. The disturbance takes the form of a bubble or cloud around the rocket and a wake of some extension behind the rocket. To minimize this disturbance, the propellant tanks on rocket NN3.08F were sealed at the end of the propulsion period (that is, at a rocket altitude of 46 km).

Since in the measurement of collision frequencies the results from a descending rocket (NN3.07) are in agreement with those of an ascending rocket (NN3.08F), it is felt that the rocket wake does not contribute a serious error to the experiment.

The bubble enclosing the rocket is affected by rocket out-gassing, a negative DC rocket potential, and RF field of the transmitting antenna.

1. *Out-gassing*—The effects of out-gassing can be dismissed on the basis of the results obtained from rocket experiments [*Horowitz* and *LaGow*, 1957] that measure atmospheric pressures and densities. In those experiments ionization gages mounted on the surface of the rocket read the ambient pressures up to an altitude of 90 km for either sealed or unsealed rockets. It is only above 90 km that rocket out-gassing becomes important, and there a different technique must be employed to determine the ambient pressure.

2. *Rocket potential*—Assuming that the photoelectric effect can be ignored, the rocket skin will, in the ionosphere, acquire a net negative charge due to the difference between the electron and positive ion velocities. *Jastrow* and *Pearce* [1957] have treated the case of an earth satellite above the *F*-region of the ionosphere. The effect of a surface charge is to create an electric field which alters the electron density in the medium around the rocket. The rocket potential ϕ is determined by the condition that the positive ion flux to the rocket surface is equal to the flux of electrons with energies greater than $e\phi$. The calculation shows that for the *D*- and lower *E*-region of the ionosphere the rocket will acquire a potential of the order of a few tenths of a volt.

In addition to the above process, the rocket acquires a potential due to the operation of the transmitting antenna, the two arms of which, when immersed in an ionized medium, act as a full wave rectifier. Experimental evidence obtained by another rocket group at this Laboratory indicates that this effect produces, in the *D*- and lower *E*-region of the ionosphere, a negative rocket potential of the order of 10 volts (J. F. Clark, private communication). Such a potential would produce an electron density distribution that is zero at the rocket surface and rises to its ambient value in a distance of the order of one meter. Since this is a very small distance compared with the total path length between the rocket and ground station, over which path the difference absorption integral is measured, it cannot appreciably affect the measurement of $\ln (E_0/E_z)$.

3. *RF field of the antenna*—Under optimum conditions the rocket antenna (half length of 4.3 meters) operates at 200 volts of RF and radiates a total power of 2 watts. At distances beyond 6 meters (one wavelength divided by 2π) the electron temperature is only a few per cent above thermal. At very small distances from the antenna (for example, $\simeq$20 cm) the RF field is sufficient to produce ionization under the proper atmospheric pressure conditions. Impedance measurements [*Jackson* and *Pickar*, 1957]

made during the rocket flight indicate that in the 60 to 105 km altitude region the antenna undergoes a detuning process. Laboratory experiments simulating flight conditions were performed to investigate this effect. It was found that the detuning is caused by a glow discharge along the length of the antenna.

The ionization formed in this discharge produces a systematic error in the measurement of the ambient electron density. This error must be negligibly small since no discontinuities in the beat note frequencies are observed on the flight record when the glow abruptly extinguishes at 105 km. In the laboratory experiments on the antenna it was found that when the glow abruptly extinguished at a pressure corresponding to 105 km, the ionization from the glow, as collected on a probe, would also disappear.

The ionization formed in the glow discharge also produces a systematic error in the measurement of $\ln (E_0/E_x)$. However, this error must be negligibly small for two reasons. First, the dimensions of the glow discharge are small compared with the path length over which $\ln (E_0/E_x)$ is measured, and second, the difference absorption $\kappa_x - \kappa_0$ goes to zero when the electron collision frequency becomes very large as it does in the glow discharge.

C. *Distribution of power between the two polarization modes*

Another possible source of error in the present experiment is the neglect of the term

$$\frac{d}{dz} \ln \frac{E_0'}{E_x'}$$

in (12). The terms E_0' and E_x' refer to the magnitudes of the radiated field strengths as they leave the antenna. In general E_0' is not equal to E_x', since in the ionosphere, each polarization mode presents a different radiation resistance to the antenna.

The magnitude of this effect can be estimated by measuring the value of

$$\frac{d}{dz} \ln \frac{E_0}{E_x}$$

under the following conditions:

1. No absorption is present between the transmitting and receiving antennas.
2. The transmitting antenna is in a medium of ionization content similar to that whic was present during the experiment.

These conditions were obtained during anothe rocket flight, NN3.11F, made on the night o February 4, 1958. From the record of that fligh it is estimated that the neglect of the term

$$\frac{d}{dz} \ln \frac{E_0'}{E_x'}$$

in (12) introduces a systematic error of abou -0.01 into the value of $\kappa_x - \kappa_0$. The minus sig indicates that correcting for this source of erro decreases the values of the experimental collisio frequencies, the magnitude of the decrease bein less than 10 pct in the center and less than 20 pc for either extreme of the altitude region o Figure 4. Here again the systematic error i small compared with the random error presen on the experimental points, and is therefor ignored.

V. Discussion and conclusion

The theoretical expression for the electro collision frequency in terms of the fundamenta parameters of temperature T, neutral particl density N, and the collision cross section [*Nicolet*, 1953], is

$$\nu = \frac{4}{3} \left(\frac{8kT}{\pi m_e} \right)^{\frac{1}{2}} N\sigma \qquad (14$$

The values of N and T used in the computation of this expression refer to a latitude of 32°N and for that latitude are very difficult to cha lenge. The present measurement of collisio frequency refers to a latitude of 58°N. Ver preliminary results of the IGY program [*Strou* 1958] give some indication of a latitude depend ence for the density in the 60 to 80 km altitud region, the density decreasing with increasin latitude. However, the Arctic density measure ments of *LaGow* and others, [1958], did no reveal any latitude effects. The possibility tha a latitude effect exists which is large enough t explain the present disagreement by a factor o three in collision frequency, although remot cannot be completely excluded at this time.

The present discrepancy between the measure and theoretical values of the collision frequencie might be explained by the value of the collisio cross section used in Nicolet's calculation. Th

calculation is based on the value of σ determined in a laboratory experiment by *Anderson* and *Goldstein* [1956] who report a value of $\sigma = 17 \times 10^{-16}$ cm^2 for 0.04 ev electrons on nitrogen. This value is equal to 60 atomic units which is larger by a factor of four than an earlier determination by Phelps and others [1951].

In view of these possible uncertainties in the parameters required for the computation and the absence of gross systematic errors in the NRL measurement, we conclude that the electron collision frequencies in the Arctic for the 60 to 80 km altitude interval are lower by a factor of three than the theoretical values of *Nicolet* [1957].

Acknowledgments—The rocket experiments were performed under the direction of J. E. Jackson to whom the author is deeply indebted for many valuable discussions and much helpful advice. The experimental techniques and methods employed were devised by J. C. Seddon. The instrumentation is due to G. H. Spaid and L. A. Lohr with assistance at the ground station by W. L. Joosten, Jr. and E. D. Lee of New Mexico College of Agriculture and Mechanic Arts.

This work was made possible by the help and cooperation of the United States National Committee for the International Geophysical Year, the Canadian Government, and the United States Army, Navy, and Air Force, all of whom made important contributions to the establishment and operation of the Fort Churchill launching facility.

References

Anderson, J. M., and L. Goldstein, Momentum transfer cross section and fractional energy loss in the collisions of slow electrons with nitrogen molecules, *Phys. Rev.*, **102**, 388–389, 1956.

Horowitz, R., and H. E. LaGow, Upper air pressure and density measurements from 90 to 220 kilometers with the Viking 7 rocket, *J. Geophys. Res.*, **62**, 57–78, 1957.

Jackson, J. E., Measurements in the *E*-layer with the Navy Viking rocket, *J. Geophys. Res.*, **59**, 377–390, 1954.

Jackson and Pickar, NRL Report 4940, 1957.

Jastrow, R., and C. A. Pearce, Atmospheric drag on the satellite, *J. Geophys. Res.*, **62**, 413–423, 1957.

LaGow, Horowitz, and Ainsworth, Arctic Atmospheric Structure, CSAGI Conference Moscow, U.S.S.R., August, 1958.

Nicolet, M., *J. Atm. Terrest. Phys.*, **3**, 200, 1953.

Nicolet, M., Pennsylvania State University, Ionosphere Research Laboratory, Internal Memo No. 171, July, 1957.

Pfister, W., The Physics of the Ionosphere—Report of Cambridge Conference, p. 394, 1954.

Phelps, A. V., O. T. Fundingsland, and S. C. Brown, Microwave determination of the probability of collision of slow electrons in gases, *Phys. Rev.*, **84**, 559–562, 1951.

Seddon, J. C., Propagation measurements in the ionosphere with the aid of rockets, *J. Geophys. Res.*, **58**, 323–335, 1953.

Stroud, W. G., U. S. Army Signal Res. & Dev. Lab., Fort Monmouth, New Jersey, May, 1958.

(Manuscript received August 16, 1958.)

Some Observations of Low-Level Ion Clouds

Charles J. Brasefield

Southern Illinois University, Carbondale, Illinois

Abstract—Simultaneous measurements have been made of the atmospheric potential at 33 meters, 21 meters, and 8 meters above ground level, using three radioactive probes. The current from each probe to ground passed through a number of 100,000 megohm resistors and the *i-r* drop across a terminal resistor of 400 megohms was measured. These measurements indicate that clouds of ions whose net charge was sometimes positive and sometimes negative frequently passed overhead at altitudes of ten meters and less. The origin of the ion clouds is uncertain, but it appears that some, if not all, were produced by motor vehicles, particularly diesel-powered vehicles.

Introduction—Very little is known about the variation with altitude of the potential gradient in the lowest hundred meters of the atmosphere. In 1933, *Scrase* [1935] made some measurements from which he concluded that "the variation of potential gradient between the ground and 10m at Kew is very small." It was decided to investigate the variation of potential gradient, under conditions in which the air in the lowest hundred meters is likely to be stratified. For this purpose a kite balloon was used to raise a radioactive probe which was connected to an electrometer and recorder on the ground by a fine wire in which a 100,000 megohm resistor was inserted in series every four meters. When these experiments were first undertaken, readings of atmospheric potential were taken at altitude intervals of four meters as the balloon was raised to a maximum altitude near 100 meters, and the readings were repeated as the balloon was lowered. Usually it was found that the atmospheric potential increased rather uniformly with altitude at a rate of about 200 volts per meter. Occasionally, abnormally high potentials were observed at certain altitudes, but after a few minutes the potential would usually fall to a normal value. In order to determine whether these sporadic increases in potential occurred at all altitudes or were limited to one altitude, it was decided to make simultaneous measurements of potential at three altitudes. Data thus obtained indicate that occasionally the potential at one altitude will rise briefly; the change in potential at the other two points may be relatively small. Such a rise in potential was found frequently at an altitude of eight meters, suggesting that a cloud of ions whose net charge was sometimes positive and sometimes negative had passed overhead at low altitude.

Experimental Equipment—A kite balloon was used to raise two radioactive probes to altitudes of 33 meters and 21 meters, respectively, and a long bamboo fishing pole was used to raise a third probe to an altitude of eight meters. Each radioactive probe consisted of a brass disc about one cm in diameter, having both faces coated with polonium. During the period of the experiments, the probe at 33 meters had a strength of about 0.2 millicurie; the other probes were about one millicurie. The probe at 33 meters was supported ten meters below the balloon by a line consisting of four meters of 12-mm garter elastic, a rod of Teflon five cm long (to insulate the probe from the balloon and mooring line), and a six-meter length of nylon cord. The probe at 21 meters was supported 22 meters below the balloon by a similar line. Apparently, field distortion due to the balloon was not serious, because the potential gradient was found to be reasonably constant with height except when ion clouds were present (see Fig. 3).

Each of the three probes was connected to a meter on the ground by means of a conductor designed to minimize the possibility of a corona discharge. The conductor consisted of four-meter lengths of fine wire joined by 25-mm Teflon insulators. Each insulator was shunted by two Victoreen Hi-Meg resistors in series, each of value 1×10^{11} ohms, accurate to one pct. The total resistances (nominal) of the 33, 21, and 8 meter lines were 15×10^{11} ohms, 9×10^{11} ohms and 2×10^{11} ohms, respectively. The lower end of each conductor was connected to the input terminal of a Keithley Electrometer

(Model 210), whose output was fed into an Esterline Angus recorder. The input terminal of the electrometer was connected to ground through a 4×10^8 ohm resistor with a 10^{-8} farad capacitor in parallel, so the time constant of the electrometer was four seconds.

The electrometers read directly in volts (generally the 0–8 volt range was used). Since the total resistance of the line to the 33-meter probe was 15×10^{11} ohms, the potential of this probe with respect to ground is approximately $15 \times 10^{11}/4 \times 10^8$ or 3750 times the reading of the electrometer. This multiplying factor is subject to correction (a) for the resistance of the probe itself and (b) because the resistance of a Victoreen resistor decreases as the potential difference across its terminals increases (at 1000 volts the resistance is about half nominal value). The resistance of the probe at 33 meters was found to be 8×10^{10} ohms; each of the other probes had a resistance of about 5×10^{10} ohms. The resistance of the probes depends on ventilation and may be expected to vary as much as 20 pct over the range of wind speeds used in these experiments [*Wichmann*, 1952]. The maximum error in potential measurements from this source was about four pct.

To determine whether the potential of the probe supported by the fishing pole eight meters above the ground was influenced by the proximity of the pole at ground potential, the balloon was used to support another probe at an altitude of eight meters. On the day when this experiment was undertaken, the potential at eight meters was quite steady. The potential indicated by the balloon-supported probe, averaged over a four minute interval, was 1380 volts. The mean potential of the probe on the fishing pole over the same interval was 960 volts. The ratio of these two values is 1.44, and all subsequent measurements of potential using the probe on the fishing pole were multiplied by this factor. The validity of this factor was confirmed by the consistency of subsequent data (see Fig. 3, for example).

Procedure—The number of days on which field experiments could be run was seriously limited, because the behavior of the balloon was unsatisfactory both at high and at very low wind speeds. If the wind exceeded ten knots, the balloon was unstable and the line of resistors was shaken vigorously, particularly if there was much turbulence. When the wind dropped much below five knots, the balloon had barely enough lift to keep the lines erect. No attempt was made to fly the balloon during periods of precipitation.

The experiments were carried out in the center of a baseball field so that the nearest poles and trees were about 65 meters away. The probes at 33 m, 21 m, and 8 m were each connected to a station on the ground consisting of a Keithley Electrometer and an Esterline Angus recorder, previously described. These stations were about 6.5 meters apart along a straight line which was perpendicular to the direction of the wind. The wind speed was estimated by observing the behavior of the balloon. These estimates were occasionally checked with the anemometer at an airport five km distant. Consequently, the wind speeds reported could easily have been in error by 25 pct but are not likely to be off as much as 50 pct.

At the beginning of each run, the three recorders were synchronized; that is, the charts were adjusted so that the pens simultaneously crossed a horizontal line. Every fifteen minutes during a run, the zero reading of each electrometer was checked by shorting the input terminal to ground, and necessary adjustment was made. At this time the synchronization of the recorders was also checked, and adjusted if necessary. After zeroing, each electrometer returned to equilibrium in about 20 seconds. It should also be noted that after an ion cloud passed, each electrometer returned to normal in about 20 seconds, indicating that the distributed capacitance of the lines to the probes must have been very small.

Results—Although the experiments were started during March, 1956, the first satisfactory data were not obtained until February 21, 1957. Some of the more interesting results obtained since then are shown in Table 1. These data have been selected from data taken on 20 separate days over a period of more than five months, the average daily record having a duration of two hours. From these data, the mean potential gradient of the layers bounded by the probes was determined. Values of the mean gradient, plotted at the center of the layers, are shown in Figures 1 through 7. The measurements of gradient are believed to be reliable to about 10 v/m.

Figure 3 illustrates an interesting situation

TABLE 1—*Selected data on potential gradient*

Date	Weather	Wind	90th Meridian Time	V_{33} (volts)	V_{21} (volts)	V_8 (volts)
			h m			
Mar 20, 1957	Clear	NW6K (Northwest 6 knots)	10 47	9600	6100	2190
			10 51.5	11,350	7300	3060
			10 54	9800	6250	2340
			14 42	7800	4700	2020
			14 43	7550	4300	860
			14 44	8200	5250	2190
Mar 21, 1957	Overcast with stratus clouds	E5K	9 42	7550	4400	1690
			10 04	2100	1300	430
			10 17	5650	3600	1320
			10 36	400	400	120
			10 44	10,150	6600	2500
			10 47	15,300	9900	3840
			10 48	Rain and sleet started		
Apr 28, 1957	Sun visible through stratus layer. (One hour earlier, light fog and some light drizzle)	NE5K	9 33.5	12,000	7700	3190
			9 34.75	11,150	7000	2340
			9 35.75	12,000	7600	3060
			9 40.5	10,300	6500	2590
			9 41.5	10,300	6350	1830
			9 42.25	10,700	6800	2620
June 5, 1957	Broken alto stratus	NE2K	9 35	8400	5900	1730
			9 35.75	8600	6000	2300
			9 37	8200	5400	2100
July 30, 1957	Broken low cumulus	N6K	9 49	3350	2300	990
			9 50.5	3050	850	840
			9 52.5	3050	2150	750

under an overcast of stratus clouds, where the potential gradient was reasonably uniform with altitude and where the magnitude of the gradient oscillated slowly between high values and low values, finally rising so that all recorders went off scale just one minute before the onset of rain and sleet. These data indicate that ion clouds of varying net charge passed overhead at altitudes greater than 33 meters. At 10h 36m a relatively intense negatively charged cloud must have passed, followed at 10h 44m by a cloud of large net positive charge which continued to increase until the end of the series of observations. Charged clouds of this sort whose altitude is small compared with their lateral dimension have been observed by *Gish* [1951].

Discussion—When satisfactory records of the potential at 33 meters, 21 meters, and 8 meters became available, it was noticed that the record at one altitude (usually 8 meters) might show a pronounced peak while the records at the other two altitudes showed little change. An outstanding case of this kind was observed on March 20, 1957 at 14h 43m (see Table 1), and was most easily explained by assuming that a cloud of negative ions had passed overhead at low altitude. Such ion clouds have been observed by others. *Whitlock* and *Chalmers* [1956] found that "the variations in the potential gradient over periods of the order of minutes were caused by the horizontal motion of wind-borne space charge contained within the first few hundred meters of atmosphere." *Holzer* [1955] reports that "a large fraction of the wind-borne space charges under stable meteorological conditions in Los Angeles lay within about 10 meters of the surface."

Undoubtedly these ion clouds are irregular in

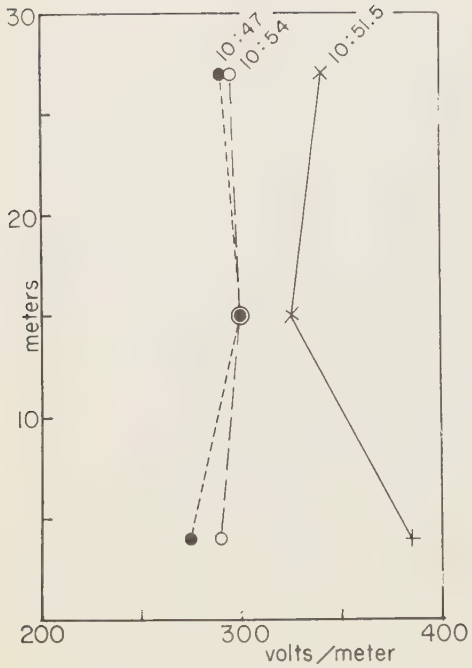


FIG. 1—Variation of potential gradient with altitude on March 20, 1957

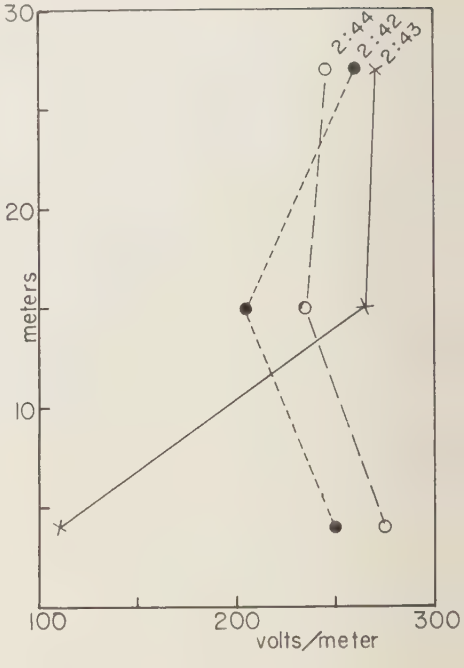


FIG. 2—Variation of potential gradient with altitude on March 20, 1957

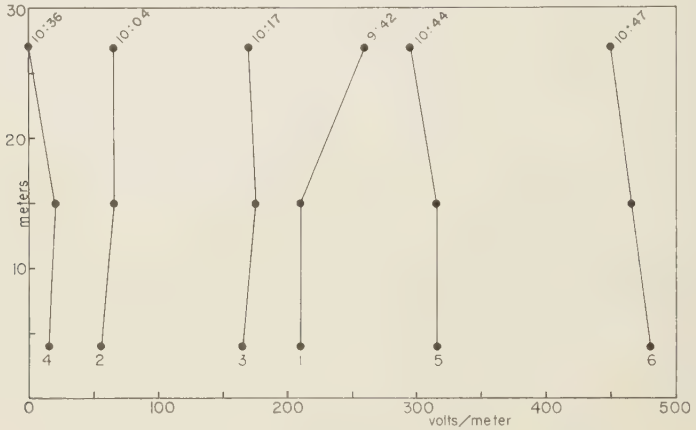


FIG. 3—Variation of potential gradient with altitude on March 21, 1957

shape and of complex charge distribution. However, if the simplifying assumption is made that an ion cloud has the shape of an infinitely long horizontal cylinder whose axis is at a known height above the earth's surface, the cylinder being filled with a charge of uniform volume density, then it is possible to compute the electric potential at altitudes of 8 m, 21 m, and 33 m in a vertical plane through the cylinder. The mean potential gradient for these layers can then be computed. The results of such calculations are shown in Figures 8, 9, and 10 for cylinder

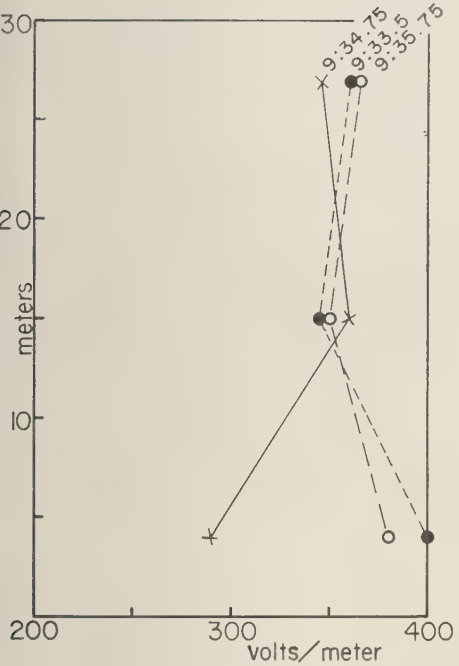


FIG. 4—Variation of potential gradient with altitude on April 28, 1957

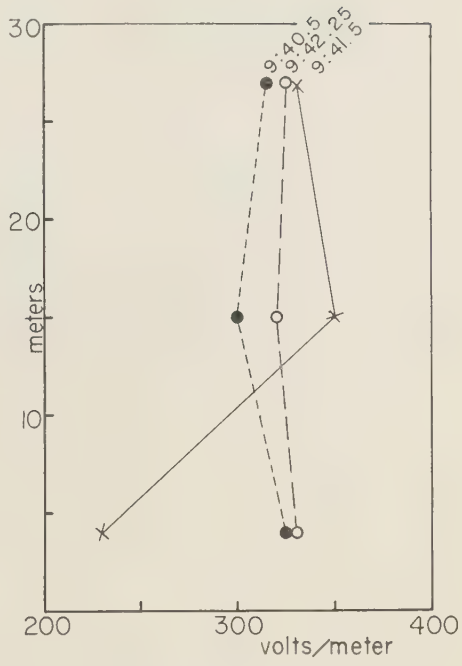


FIG. 5—Variation of potential gradient with altitude on April 28, 1957

heights of 10, 20, and 30 meters, respectively. The abscissa is the potential gradient in arbitrary units. Figure 11 indicates the type of field distortion which might be expected from a dipole.

The variation of potential gradient with altitude shown in Figure 8 is similar to that found in Figure 1 (10h 51.5m) and in Fig. 4 (9h 33.5m and 9h 35.75m). It is therefore believed that the peak observed on the record of March 20, 1957 at 10h 51.5m could have been produced by a positive ion cloud which passed overhead at an altitude of about 10 meters. This type of ion cloud will be referred to as a low-level positive cloud ($L+$). The mirror image of this configuration is found in Figure 2 (14h 43m), Figure 4 (9h 34.75m) and Figure 5 (9h 41.5m). This effect on the field is believed to be due to a low-level negative cloud ($L-$). Similarly, a configuration like that of Figure 9, due to a positive cloud at medium height ($M+$) is found in Figure 6 (9h 35.75m) while its mirror image due to $M-$ is found in Figure 7 (9h 50.5m). Very little attention was paid to a general increase in field at all altitudes, similar to that of Figure 10, because of the uncertainty in the height of the ion cloud producing it. The variation of potential gradient with altitude like that of the dipole of Figure 11 is found in Figure 6 (9h 35m) and its mirror image in Figure 2 (2h 42m).

TABLE 2—*Calculated parameters of assumed cylindrical ion clouds*

Time	h (meters)	λ (statcoulomb) (per meter)	C volts/m
March 20, 1957 (10h 51.5m)	9	25	345
March 20, 1957 (14h 43m)	8	−60	230
July 30, 1957 (9h 50.5m)	18	−65	150

It should be emphasized that the values of potential gradient plotted in Figures 1–7 are mean values over layers 12, 13, and 8 meters

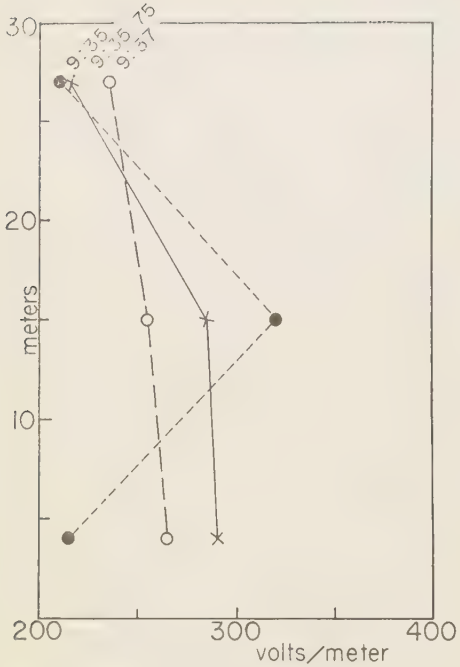


FIG. 6—Variation of potential gradient with altitude on June 5, 1957

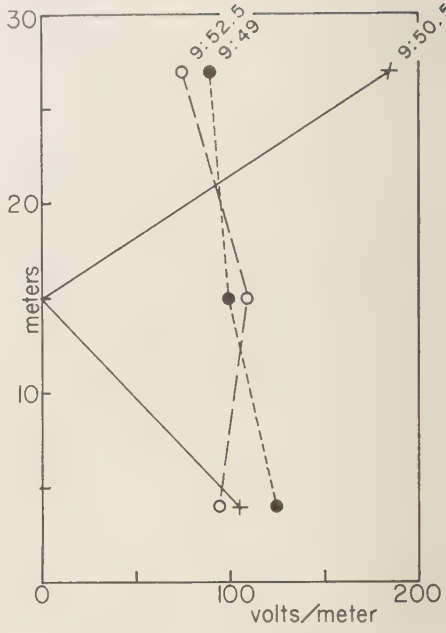


FIG. 7—Variation of potential gradient with altitud on July 30, 1957

in thickness, respectively. If it is assumed that these mean values are representative of the true potential gradient at altitudes of 27, 15, and 4 meters, it is possible to compute the height of the cylindrical cloud above the earth's surface, the charge per unit length of cylinder, and the gradient which would exist in the absence of the ion cloud. The results of three calculations of this sort are shown in Table 2.

The line charge densities shown in Table 2 are smaller than the line charge densities reported by Whitlock and Chalmers by a factor of 100 (they report an average value of 4500 statcoulombs per meter). However, the average calculated height of their line charges was 160 meters, which they believe is greater than the true height. If the height used in their calculations is reduced, their line charge densities would be reduced proportionally.

Whenever the change in potential gradient from 4 to 15 meters, or from 15 to 27 meters, was equal to or greater than 30 volts per meter, it was assumed that the distortion of the field was due to an ion cloud in the vicinity, and an attempt was made to determine the sign and height of the cloud. A summary of the results, for the perio February 21, 1957 to July 30, 1957, is shown i Table 3.

TABLE 3—*Statistics of ion clouds*

Type of ion cloud	Occurrences
$L+$	36
$L-$	52
$M+$	41
$M-$	30
Dipole (+ above −)	11
Dipole (− above +)	10

The question naturally arises as to the origi of these ion clouds, particularly whether or n they are man-made. Common sources of io clouds mentioned in the literature include exhau gases from automobiles, locomotives, and chim neys, brush discharges from high voltage line and dust kicked up from the earth's surfac [*Mühleisen*, 1956]. If the ion clouds listed i Table 3 originated in a source such as one

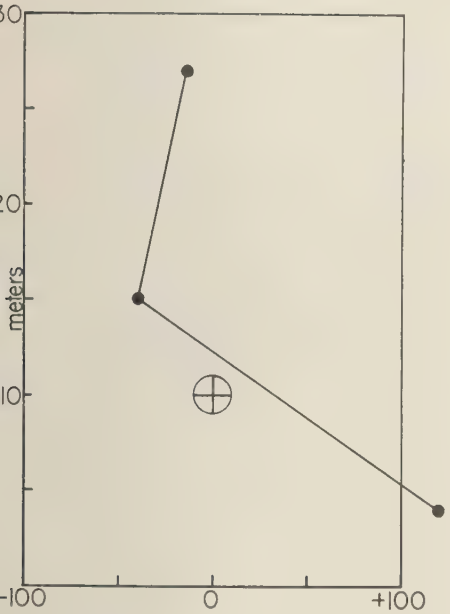


FIG. 8—Variation of potential gradient with altitude above and below an infinite cylinder of positive charge located 10 meters above the ground

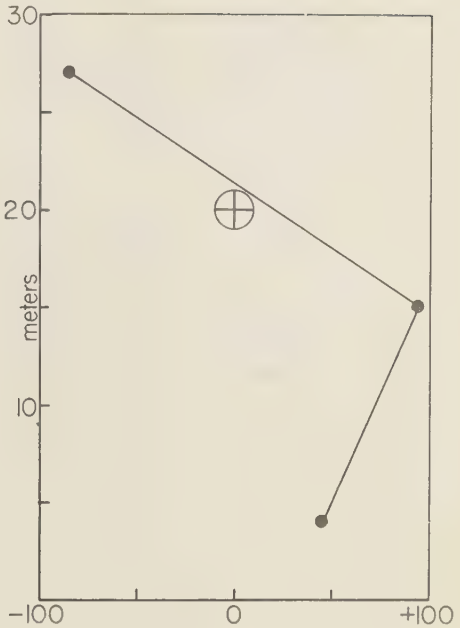


FIG. 9—Variation of potential gradient with altitude above and below an infinite cylinder of positive charge located 20 meters above the ground

hese, it might be expected that the prevalence of he ion clouds would depend on the wind direc- on. No such relationship was obvious. Auto- nobiles passing upwind from the equipment sually had no effect on the records. There was ome evidence that the ion clouds were more revalent, (a) when the campus was active, b) when the potential gradient was high.

On April 16, 1958, during a routine experiment n the baseball field, it so happened that two rge earth-movers with diesel engines began naking periodic trips back and forth along a oad which ran along the north side of the ball eld, about 300 feet from the probes. The wind as light and variable, ranging from northwest o east. When an earth-mover passed and the ind was blowing toward the equipment, very narked increases in atmospheric potential were ecorded at all three altitudes after a time interval onsistent with the wind velocity. Only about ne out of ten passenger cars which passed during is period produced a noticeable effect on the ecords of potential; when perceptible, the effect as positive and was much smaller than the ffect produced by a diesel-powered earth-mover. n analysis of data for one of the clouds from an earth-mover yielded the following constants: $h = 19$ meters, $\lambda = 75$ statcoulombs per meter, $C = 260$ volts per meter (this is about twice the normal field, so there must also have been positive ions at altitudes above 30 meters). It thus appeared that a diesel engine might be a rich source of positive ions.

To test this conclusion, a space charge detector was made in the form of a cylindrical Faraday cage of window screening, 50 cm in diameter and 50 cm long. In the center of the grounded cage was placed a polonium button which was connected directly to the input terminal of a Keithley electrometer on which the cage rested. Such a cage is quite sensitive to space charge [*Vonnegut* and *Moore*, 1958], as can be shown by placing it downwind from a source of positive ions such as the flame of a plumber's soldering torch which is 100 volts positive with respect to ground. This cage was used to test the exhaust gases of the diesel engine of an earth-mover, and they were found to be strongly charged, positive. The exhaust gases of passenger automobiles were found to be sometimes positive, sometimes negative, and occasionally uncharged. The exhaust gases of one diesel locomotive tested

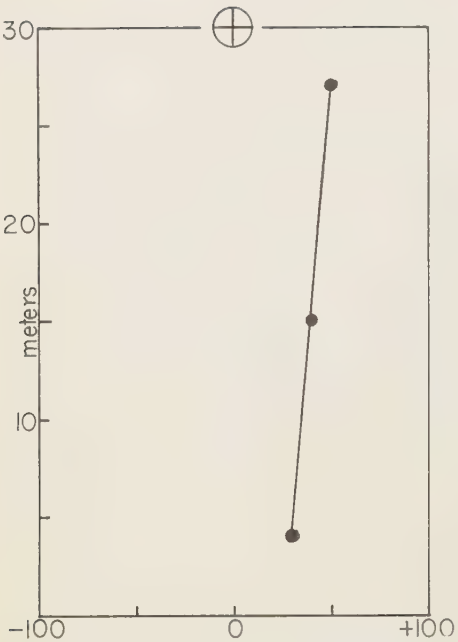


FIG. 10—Variation of potential gradient with altitude above and below an infinite cylinder of positive charge located 30 meters above the ground

proved to be uncharged. An attempt to test the exhaust of an airplane was unsuccessful because of the propellor blast; however, it has been shown [*Stimmel* and others, 1946] that under certain conditions, ionization of the exhaust gases plays an important role in discharging electricity from an airplane. It is not clear why the exhaust of some motor vehicles should be charged, but it appears that some if not all of the ion clouds detected in this series of experiments were produced by motor vehicles, particularly diesel-powered vehicles.

The ease with which ion clouds can be produced and detected suggests that they would be useful in tracing the trajectories of air parcels.

Acknowledgments—This experiment was suggested by H. J. aufm Kampe and H. W. Kasemir of the Signal Corps Laboratories, both of whom have been most generous with advice during the course of the work. It is a pleasure to acknowledge also the assistance of the following students in the field experiments: Paul Marchionda, Robert Montgomery and Robert Revak. Our laboratory mechanic, Robert Griggs, designed and constructed much of the field equipment. Thanks are also due to O. H. Gish of Fort Pierce, Florida, for suggesting improvements in this paper, including particularly the possibility of making a detailed analysis of the ion clouds.

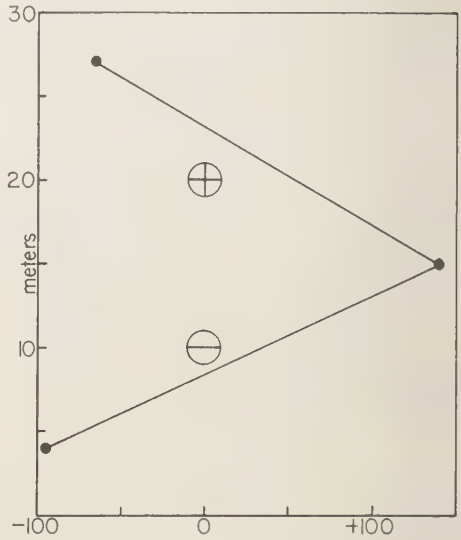


FIG. 11—Variation of potential gradient with altitude above and below a dipole consisting of an infinite cylinder of positive charge at 20 meters and an equal negative cylindrical charge at 10 meters

REFERENCES

GISH, O. H., The universal aspects of atmospheric electricity, in T. F. MALONE, ed., *Compendium of Meteorology*, Amer. Met. Soc., Boston, 1951, 101–119, esp. p. 110.

HOLZER, R. E., *Studies of the universal aspect of atmospheric electricity*, Univ. Calif. Inst. Geophys. 84 pp., 1955 [photocopies available as listed under PB 119024 in *U. S. Govt. Res. Rep.*, **25**, 65, 1956] esp. p. 11.

MÜHLEISEN, R., On the deviations of the course of elements of atmospheric electricity on continents from the worldwide course, *J. Atmos. Terr. Phys.* **8**, 146–157, 1956, esp. p. 151.

SCRASE, F. J., *Some measurements of the variation of potential gradient with height near the ground at Kew Observatory*, London, Met. Off., Geophys. Mem. No. 67, 1935.

STIMMEL, R. G., E. H. ROGERS, F. E. WATERFALL AND R. GUNN, Electrification of aircraft flying in precipitation areas, *Proc. Inst. Radio Eng.*, **34**, 167–177, 1946, esp. p. 175.

VONNEGUT, B., AND C. B. MOORE, *A study of techniques for measuring the concentration of space charge in the lower atmosphere*, Final Rep., p. 48, 1958.

WHITLOCK, W. S., AND J. A. CHALMERS, Short period variations in the atmospheric electric potential gradient, *Q. J. Roy. Met. Soc.*, **82**, 325–336, 1956.

WICHMANN, H., Die Messung des luftelektrischen Feldes mit radioactiven Potentialsonden, *Arch. Met.*, **5**, 90, 1952.

(Manuscript received October 10, 1958.)

Photometric Observations of the 5577 A and 6300 A Emissions Made during the Aurora of February 10-11, 1958

E. R. MANRING AND H. B. PETTIT

Geophysics Research Directorate
Air Force Cambridge Research Center
Air Research and Development Command
L. G. Hanscom Field, Bedford, Massachusetts

Abstract—The intense auroral activity of February 10–11, 1958, was recorded on airglow photometers at Sacramento Peak, New Mexico, and at Tonantzintla, Mexico. At Sacramento Peak the whole sky varied in intensity from 10^5 to 10^8 Rayleighs for the 6300 A emission, and from slightly less than 10^3 to more than 10^4 Rayleighs for the 5577 A emission. The ratio of the intensity of the 6300 A emission to that at 5577 A was about 2.5×10^3 throughout the night. There was no enhancement of the 5890 A emission. At Tonantzintla, the 5577 A emission varied from 200 to 600 Rayleighs. Contour lines of equal intensity were nearly perpendicular to the geomagnetic meridian at both sites. An interesting correlation occurs between absorption measurements of cosmic radio noise made at Boulder, Colorado, and the intensity variation of the 5577 A and 6300 A emissions as observed at both Sacramento Peak and Tonantzintla.

Introduction—The airglow photometer at Sacramento Peak, New Mexico, 32°47′N and 105°49′W, has been used for regular observations of the airglow emissions at 5577 A, 5890 A, and 6300 A on clear nights since February, 1955 [*Manring and Pettit*, 1958]. During this period seven nights showed definite auroral characteristics. Since the aurora of February 10–11, 1958, was by far the most intense, this aurora is the only one discussed here. It was visible throughout the entire southwest (see *Sky and Telescope*, vol. 17, pp. 280–283, 1958) and passed south of the zenith of Sacramento Peak. This aurora also appeared in the observations of the 5577 A emission being made simultaneously at Tonantzintla, Mexico, 19°02′N and 98°19′W. The instrument at Tonantzintla is not equipped to measure the emissions at 6300 A and 5890 A.

Observations—The airglow photometer at Sacramento Peak has been described elsewhere [*Dunn and Manring*, 1956]. A sky survey consists of a zenith intensity reading and continuous scans in the azimuth at zenith distances of 40, 55, 70, 75, and 80 degrees. Each survey, for which a single emission is recorded, requires 5m 50s. The field of view of the photometer is 5°. Since the 6300 A emission was the most prominent, more surveys were taken of it, with 5577 A surveys interspersed, and only occasionally were surveys made in 5890 A. Because of the necessary sporadic nature of the observations and the relatively large field of view, only the large scale motions and intensity fluctuations could be recorded.

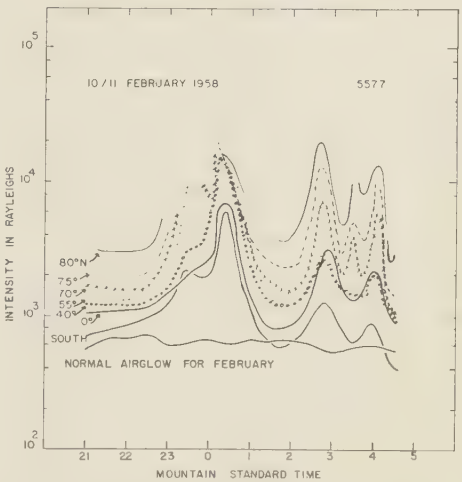


FIG. 1—The diurnal variation of the 5577 A emission for February 10–11, 1958, as seen from Sacramento Peak, New Mexico. Seven of the plots represent the diurnal variation as observed for the following parts of the sky: azimuth north and zenith distances of 80, 75, 70, 55, and 40 degrees; the zenith; and azimuth south at zenith distance 75°. The normal airglow intensity is also shown. Between 2300 and 0030 the aurora was so bright that the recorder went off scale for north 80, 75, and 70 degrees

The maximum intensity of the 6300 A emission

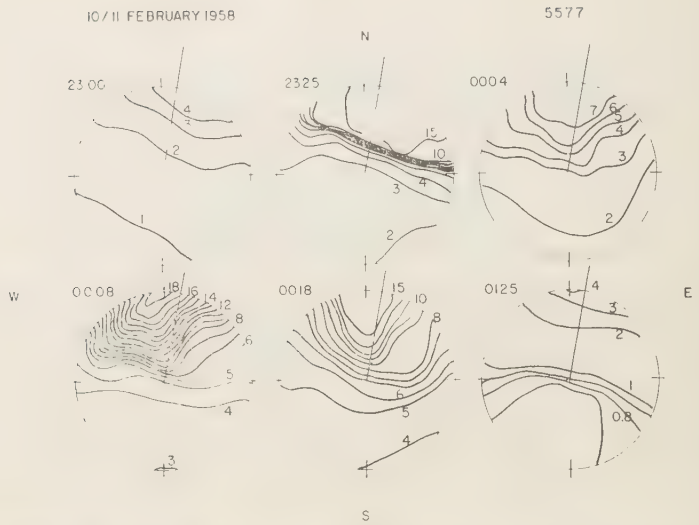


FIG. 2.—A detailed study of circular plots for the 5577 A emission for February 10–11, 1958. The times are in MST and indicate the start of the survey. The isophotes are in Kilorayleighs. The direction of the magnetic field is indicated

was 10^8 Rayleighs (see *Transactions of the International Astronomical Union*, vol. IX, edited by P. Th. Ooterhoff, p. 318, Cambridge University Press, 1957), and the 5577 A intensity was a little over 10^4 Rayleighs. The sodium emission at 5890 A was also observed and showed no detectable enhancement. Figures 1 through 4 show a series of circular plots and time variation curves of the maximum intensity occurring in respective azimuth scans for the 6300 A and 5577 A emissions. These circular plots have been drawn with an assumed layer of intensity at 100 km to establish a scale. This probably does not bear a close relation to geographical intensities since the auroral 6300 A emission was probably much higher than 100 km and there is no reason to believe that it was confined to a layer.

Analysis of the data—The following parameters can be determined from an analysis of the above data: (1) time variation in intensity of the aurora; (2) direction of maximum intensity; (3) comparison of the intensity of the 6300 A emission to that at 5577 A; and (4) large scale movements. A discussion of these points follows:

(1) Three maxima were observed in both the 5577 A and 6300 A emissions. The first maximum was the brightest and occurred at 2345 MST. Two other minor maxima occurred at 0200 and

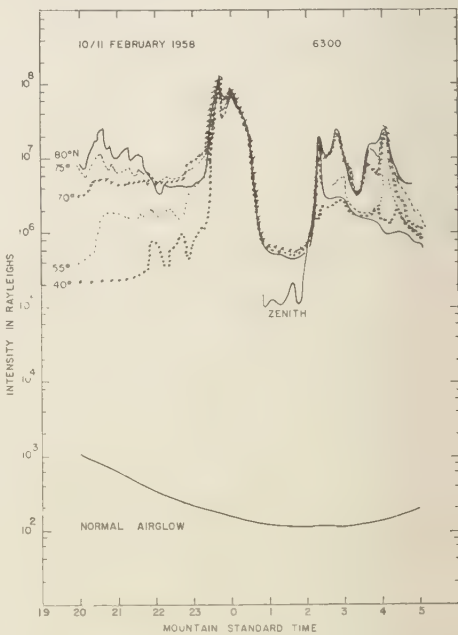


FIG. 3—The diurnal variation of the 6300 A emission for February 10–11, 1958, as seen from Sacramento Peak, New Mexico. The method of plotting is the same as that for Fig. 1

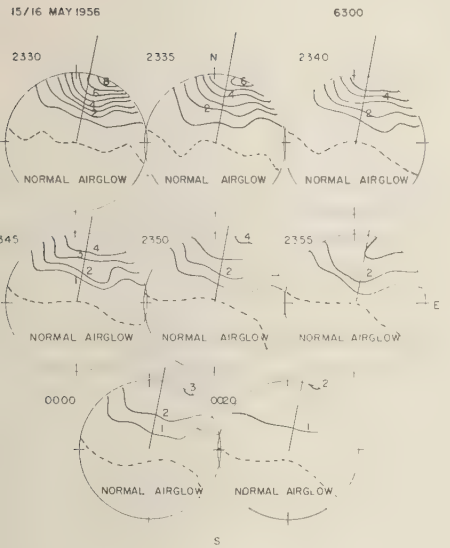


FIG. 4—A detailed study of circular plots from ie 6300 A emission for February 10–11, 1958. The mes are in MST and indicate the start of the rvey. The isophotes are in Rayleighs. The direc- on of the magnetic field is indicated

400 MST. This pattern of events agrees well ith an analysis of systematic observations made om various auroral stations reported by *Vitra* [1952].

(2) The direction of maximum intensity was bserved for each azimuth scan. There was no ppreciable change in this direction with time, s illustrated in Figures 2 and 4. The average irection for the 6300 A emission is N21.6°E. he deviation around this mean was only about ±11°. For the 5577 A emission the direction was 20°E for the first part of the night, changing north during the rest of the night. Since the irection of the magnetic field for Sacramento eak is N12.5°E, the contour lines of equal tensity for both emissions are nearly perpen- icular to the geomagnetic meridian.

(3) Since observations of the 5577 A and 300 A emissions were not made simultaneously, was necessary to compare adjacent surveys. or example, an intensity ratio of the 6300 A mission to that at 5577 A was obtained by using ie intensity of the 6300 A emission for a selected oint in the survey starting at 0230 MST, and ie intensity of the same point from the survey the 5577 A emission starting at 0235 MST. Eighty such ratios were obtained from the selected adjacent surveys as given in Table 1. In order to determine the observational scatter of these ratios, the intensities of the 5577 A emission to those at 6300 A for the eighty points were plotted as shown in Figure 5. The average line was drawn by eye since the observations do not warrant a statistical treatment. This gives

TABLE 1—*Starting times for the adjacent surveys used for obtaining intensity ratios of the* 6300 A *and* 5577 A *emissions*

5577 A	6300 A
0235 MST	0230 MST
0240	0245
0250	0245
0250	0255
0328	0334
0344	0350

a ratio of the intensity of the 6300 A emission to that at 5577 A of about 2.5×10^3. The scatter appears to be random with respect to azimuth and zenith distance, but is more pronounced for

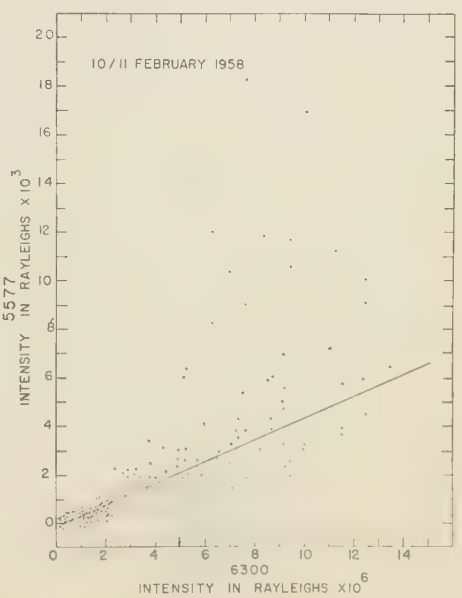


FIG. 5—A comparison of the intensity at 6300 A to that at 5577 A for February 10–11, 1958

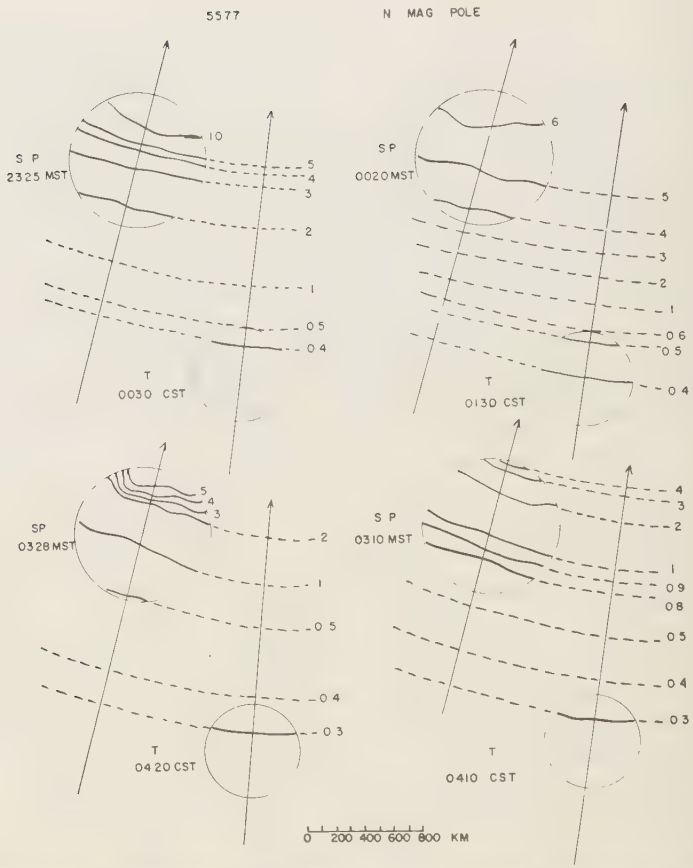


FIG. 6—A geographical representation of the simultaneous observations made of the 5577 A emissio from Sacramento Peak (SP) and Tonantzintla (T). The isophotes are in Kilorayleighs

larger intensities. Points falling below the mark of zero intensity represent the normal airglow.

(4) Figures 2 and 4 show a definite large scale apparent motion in both auroral emissions. Since the edge of the 6300 A enhancement was pronounced, it was possible to follow the progression of this edge from the NNE part of the sky through the zenith to the SSW horizon. The position of this edge was plotted against time for 40, 55, 70, and 80 degrees zenith distance. The edge was straight within the limits of observation and remained perpendicular to the geomagnetic meridian to within $\pm 5°$. The velocity of such a motion cannot be obtained from a single observing station unless the height to the emission is known. If a constant height is assumed, then by simple geometry the rati of the north-south velocity to this height i roughly constant and equal to about 0.7 whe the unit of time is in hours. This value is give as a measured parameter with no physica significance attached to it since the heigh is unknown. It was possible to follow the Kilorayleigh contour line for the 5577 A emissio from the northern to southern horizon and back with a similar ratio of velocity to height of 6 fo both north to south and south to north apparen motion.

Simultaneous photometric observations of th 5577 A emission were also made at Tonantzintla Mexico. The maximum intensity there wa about 600 Rayleighs occurring at 0030 CST

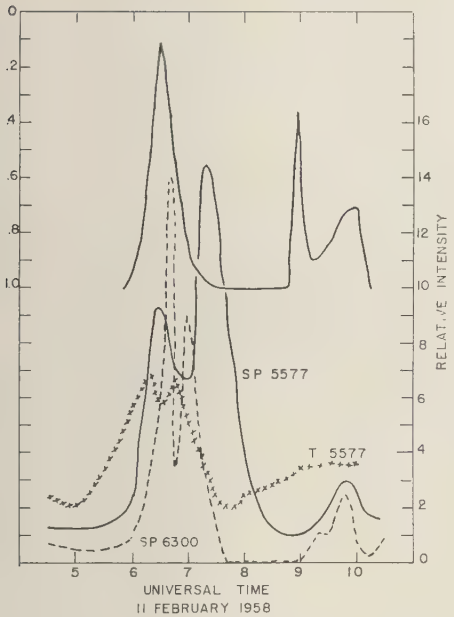


FIG. 7—A composite plot of the diurnal variation f the relative intensity of the 5577 A and 6300 A missions for the whole sky appears in the lower ortion of this figure. The curves labeled SP refer o the Sacramento Peak observations and the urve labeled T refers to the Tonantzintla observaions. The upper portion of the figure represents he absorption of the cosmic radio noise

Because of the irregular period of observation f 5577 A from Sacramento Peak, no definite omparison of time of maximum can be made. However, it appears that the maximum occurred almost simultaneously at both places. Figure 6 shows a series of simultaneous plots obtained from the two stations and plotted as described above. Again, contour lines are nearly perpendicular to the direction of the geomagnetic pole [*Mitra*, 1952].

During this night of February 10–11, 1958, the absorption of cosmic radio noise [*Warwick*, 1958] was being observed at the High Altitude Observatory in Boulder, Colorado. Figure 7 is a composite plot of the relative diurnal variation of intensity of the 5577 A and 6300 A emissions for the whole sky, and the absorption of cosmic radio noise. At the times when the absorption of cosmic radio noise is greatest there is a definite correlation with the intensity maxima for both 5577 A and 6300 A at all points of the sky over Sacramento Peak and Tonantzintla. It is remarkable that sharp changes over wide areas and of different parameters should show such simultaneity.

REFERENCES

DUNN, R., AND E. R. MANRING, Recording night sky photometer of high spectral purity, *J. Opt. Soc. Amer.*, **46**, 572–577, 1956.

MANRING, E. R., AND H. B. PETTIT, The 5577 A emission of [OI] in the night airglow from Sacramento Peak, New Mexico, *J. Geophys. Res.*, **63**, 39–49, 1958.

MITRA, S. K., *The upper atmosphere*, second edition, The Asiatic Society, Calcutta, pp. 413 and 418, 1952.

WARWICK, JAMES W., Absorption of cosmic radio noise during the great aurora of 11 February 1958, *Science*, **127**, 1047–1048, 1958.

(Manuscript received October 6, 1958.)

Progress in Cosmic Ray Research since 1947

B. PETERS

Tata Institute of Fundamental Research
Bombay, India
and University Institute of Theoretical Physics
Copenhagen, Denmark

Abstract—Cosmic ray physics, which ten years ago was a fairly specialized branch of science, has in the course of its recent development become closely linked with many other fields of research. It has become an integral part of astrophysics, radioastronomy, and solar physics; it has made important contributions to such diverse fields as geomagnetism, hydrology, and archeology, and has begun to gain some importance in the study of meteorites and of oceanography and meteorology. It has also given rise to one of the newest and most active branches of physics, 'particle physics,' and has thereby provided the impetus for designing the large and powerful particle accelerators which are now in operation or under construction in various parts of the world. This paper is a survey of the most important developments which have taken place in cosmic ray physics during the last decade.

The principal technical advances which have exerted decisive influence on cosmic ray research are the development of large stratosphere balloons, of nuclear emulsions, and of radio telescopes, as well as the great progress made since the war in radio chemistry and in low level beta- and gamma-ray counting.

The three sections following the introduction discuss how, with the help of these and other techniques, the nature of the primary cosmic ray particles was determined and how a connection was established between cosmic ray effects and astrophysical and solar phenomena. The next section deals with largely successful efforts which have been made to understand the complicated chain of processes by which the energy carried by primary cosmic ray particles entering from outer space is first distributed over a great variety of new stable and unstable particles and finally is dissipated within the atmosphere. The two following sections treat the discovery of stable and radioactive isotopes which are produced during the passage of cosmic radiation through the atmosphere and contain a short review of the application which these isotopes have found in other branches of science. The last section contains a brief discussion of still unsolved problems which have now become accessible because of more recent technical advances such as satellites, permanent polar stations, and synoptic cosmic ray observations during the International Geophysical Year.

I. INTRODUCTION

Basic problems—Ten years ago the nature of smic radiation and its origin was largely known. Only speculative answers could be ven to such basic questions as: What are cosmic ys? Where do they come from? How are they oduced? Are cosmic rays a universal phemenon, and does their presence in distant rts of the universe give rise to observable fects on the earth? What happens when cosmic ys collide with atoms and nuclei? How do ese collisions determine the various secondary fects which cosmic rays produce in the atmosiere? Does cosmic radiation produce detectable anges in the isotopic composition of the earth? hat relation, if any, exists between cosmic ys and the evolution of life?

Today there is a large body of new factual information and it is possible to answer some of these questions, at least partially, with a certain degree of confidence. In brief outline the answers could be as follows: Cosmic rays are electrons and nuclei of various atomic numbers, which move with great velocity in the space between stars and in the region surrounding the stellar conglomerations of our and other galaxies. Their occurrence is connected with explosions in the interior or on the surface of stars. They receive their energy by interacting with magnetic fields imbedded in the ionized gas clouds which are expelled during such explosions. Their presence in distant regions can betray itself by the emission of polarized light and of radio waves. When they collide with nuclei they produce a great variety of reactions,

but most of their energy is transformed into pions (π-mesons) and other types of unstable particles. They have been incident on the earth for many millions of years and have produced detectable changes in its isotopic composition. There were probably periods of vastly increased cosmic radiation intensity in the history of the earth, and the resulting increase in the rate of mutations could have been responsible for far reaching evolutionary changes.

As far as they go these answers are probably correct, but they are not exhaustive and they raise a large number of new questions. Nevertheless, they show that the understanding of the cosmic radiation phenomenon and its relation to other fields of knowledge has progressed considerably during the last decade. It is the aim of this article to give a brief account of the main experimental discoveries and theoretical developments which have been responsible for this progress.

A brief summary of cosmic ray information available before 1947—It may be useful at this point to summarize very briefly the state of knowledge which had been reached shortly after the termination of the second world war.

It was known that the primary cosmic radiation arriving from outer space is affected by the earth's magnetic field. The variation of intensity with latitude and with angle of incidence showed that a large fraction of the primary radiation consists of positively charged particles, presumably protons. The ionization produced by cosmic rays in the atmosphere had been measured at various geographic locations with balloon-borne ionization chambers, and these measurements made it possible to construct, approximately, a primary energy spectrum for particles with energy between 3.5 and 15 Bev. It was known, however, that some primaries have much higher energies (in excess of 10^{13} or 10^{14} ev) and give rise to the so-called extensive air showers, the simultaneous arrival of large numbers of electrons spread over a wide area. Furthermore, it had been established that cosmic radiation, as measured on the surface of the earth, is practically constant in time.

As far as secondary effects in the atmosphere are concerned, it was known that cosmic ray intensity increases strongly with altitude, reaching a maximum at a height of 14 km. The strongest component in the atmosphere was identified as consisting of electrons and positron which multiply by alternating processes involvin the emission of photons by bremsstrahlung an the subsequent conversion of high energy quant into electron positron pairs. The radiation als contains unstable particles, μ-mesons, whic dominate near sea level and can penetrate t great depth below ground. The mass and lifetim of the μ-mesons had been determined. A thir cosmic ray component capable of producin nuclear disintegrations was known to exis but had not been fully identified. It was, how ever, established that it contains protons an neutrons which are themselves products o nuclear disintegrations of atmospheric nitroge and oxygen.

These are the principal results of prewar cosmi ray research. The second world war interrupte the research and brought it practically to standstill. After the war it was resumed on much larger scale than before.

New experimental techniques—Among the man technical improvements and inventions whic have characterized postwar scientific develop ment, there are four which stand out because o the decisive role which they have played i cosmic ray research.

1. Progress in the technique of *high altitud balloon flying* has made it possible to observ cosmic rays as they arrive from outer space before they are appreciably modified by collision with nuclei in the atmosphere.

2. Improved *nuclear emulsions* furnish a simpl and sensitive detector in which the charge, mass and velocity of electrically charged particles ca be measured with great precision, and in whic the process of creation, interaction, and deca of particles can be observed directly and can b studied in great detail.

3. *Radiotelescopes* have been developed whic can register very faint radio signals emitted b high energy electrons spiralling around magneti field lines in various parts of the universe. Thes telescopes can locate strong sources of radi noise with sufficient precision to establish thei association with definite astronomical objects.

4. *Radiochemical techniques* for separation an detection have become so sensitive that th minute quantities of radioactive isotopes, whic are produced by cosmic radiation, can be detecte on the earth, in the ocean, in the air, and i meteorites.

We shall now try to sketch how, by employing these and other techniques, new information on cosmic radiation was accumulated, bringing this field of research into the closest contact with many other fields of knowledge. In this article, the main emphasis will be placed on the close relation which has developed between cosmic ray research and astrophysics. But progress in other directions will also be discussed.

II. Primary cosmic radiation

What are cosmic rays?—The first direct observation of events produced by primary cosmic ray particles was made in nuclear emulsions exposed at great altitude in 1948, in one of the earliest stratosphere flights in which plastic balloons were employed to carry cosmic ray equipment. The emulsions contained tracks of particles heavier than any that had been seen before. Fission fragments had been known to produce very heavy tracks; but fragments from fission reactions come to rest after travelling a few hundredths of a millimeter in emulsions, losing all their energy by intense ionization. Many of the new particles, on the other hand, traversed a whole stack of glass-backed nuclear emulsions, corresponding to ranges of several centimeters. It was clear that they had to be attributed to massive particles of high charge with energies of many billions of electron volts (Bev = 10^9 ev).

This was proof that atomic nuclei belonging to various elements are accelerated to immense energies somewhere in the universe. Here was a new tool for investigating stellar processes, for exploring the distribution of fields and matter in the regions which the particles had to traverse before reaching the earth, and for studying the behavior of collisions between nuclei whose energies were then, and will be for many years to come, far outside the energy range to which particles can be accelerated in man-made machines.

Many questions had to be answered before these nuclei could be used as effective probes for exploring distant regions of the universe and could thereby become useful tools for astrophysical research. The most urgent experimental problems were to determine: the chemical composition of the primary radiation, the energy distribution of the various types of particles, and the intensity of the radiation; and to establish whether the particles arrive as atomic nuclei from outer space, completely stripped of their orbital electrons, or whether they arrive as atoms, molecules, or even larger aggregates of matter (dust), disintegrated and stripped only when they reach the outer fringes of the atmosphere.

These questions found answers in the two years following their discovery.

Measurement of the atomic number of a fast moving nucleus—At first sight, the problem of determining correctly the atomic number of individual nuclei belonging to a variety of elements seems formidable when viewed from the point of view of classical chemistry. However, when dealing with particles whose energies are 10^9 times larger than those customary in chemical reactions, it turns out to be comparatively simple. The specific ionization produced by particles with electric charge, Ze, and velocity, v, is nearly proportional to $(Ze/v)^2$. Nuclear emulsions can be produced whose sensitivity to ionization processes varies within wide limits. The density of the track which a particle leaves in a nuclear emulsion (for instance, the number of developed silver halide grains per unit distance) furnishes a very precise method for determining specific ionization. If the particle velocity, v, is close to that of light, such a measurement is sufficient to determine the nuclear charge Z. If the particle velocity, v, differs significantly from the velocity of light it becomes necessary to measure, in addition to ionizing power, some other parameter which characterizes the velocity. Methods for measuring particle velocity are discussed later. They are sufficiently accurate so that the atomic number of a particle can be determined uniquely, whatever its energy. The new particles were found to have atomic numbers ranging from $Z = 1$ to $Z = 26$ or 27.

The question then arises whether the particles arrive as bare nuclei, or whether they approach the earth as atoms, molecules, or dust, disintegrated and stripped of electrons only in the outer layers of the atmosphere. This was determined in the following manner.

The earth's magnetic field deflects charged particles and prevents them from reaching the atmosphere unless their momentum exceeds a certain value. This critical, or cut-off momentum depends on the charge of the particle as well as on geomagnetic latitude. With sufficient accuracy, the smallest momentum, p_c, with which a particle

can enter the atmosphere at latitude, λ, from the zenith direction is given by

$$p_c = 14.9\ Z_{eff}\ \cos^4 \lambda\ (Bev/c),$$

where Z_{eff} is the electric charge carried by each particle before entering the atmosphere. Between the 'top' of the atmosphere and a balloon-borne detector the momentum of the particle changes only by a small and calculable amount. But its effective charge may change considerably if it arrives surrounded by orbital electrons; a few $\mu g/cm^2$ of air are sufficient to strip off whatever orbital electrons it carries. The charge measured when the particle enters the emulsion is, therefore, its nuclear charge, Z. The experiments showed that the observed cut-off momentum, p_c, corresponds to the value $Z_{eff} = Z$, proving that the primary particles are atomic nuclei, not ions, molecules, or dust grains, and that they approach the earth completely stripped of orbital electrons.

Suitable detectors for various types of primary particles—The detection of heavy nuclei is simpler than that of other primary components. In emulsions their tracks stand out clearly against a large background of secondary particles because they combine long track length with very high density of ionization.

More difficult is the identification of primary electrons. One has tried to detect them by searching for fast particles in the stratosphere which carry a single electronic charge and multiply quickly when traversing a thin lead plate inside a cloud chamber. Although, as will be shown later, there can be little doubt that primary cosmic ray electrons do exist, they are rare and have not yet been detected; in a comparable energy range they are at least a hundred times less frequent than cosmic ray nuclei.

Primary protons are the most difficult particles to identify, in spite of the fact that they represent the strongest primary component. In order to distinguish them from various types of secondary particles, it is necessary to measure simultaneously their velocity and direction of motion, as well as their penetrating power and ionization. Only recently, suitable detectors consisting of Geiger counters, absorbers, scintillators, and Czerenkov counters have been constructed and used successfully for measuring the primary proton component.

Measurement of particle energy—New techniques had to be developed for measuring particle energies much higher than those which one had previously encountered in particle physics.

In the non-relativistic region, that is, below about 5×10^8 ev/nucleon, several methods are applicable. If the detector contains a sufficiently large amount of absorbing material, the particle will be slowed down appreciably by ionization losses and may be even brought to rest within the detector. Their range, or the rate at which their ionizing power changes as they traverse material, can be used to measure their energy.

If the particle energy lies in the interval from 5×10^8 to about 3×10^9 ev/nucleon, the most satisfactory energy determination comes from measuring the multiple scatterings which the particles suffer as they traverse a nuclear emulsion. One determines the deviation of the track from a straight line and this is simply related to the particle charge and momentum.

The effect of scattering becomes quite small when the particle energy exceeds a few Bev/nucleon, and the most convenient method for determining the number of cosmic ray particles in the energy interval between a few times 10^9 and about 10^{10} ev is to study their flux as a function of geomagnetic latitude. As one goes to lower and lower latitudes, the minimum energy, which a particle must have to penetrate the earth's field and enter from the zenith direction, increases. At the equator it reaches the value 1.5×10^{10} ev for protons and about 7×10^9 ev/nucleon for multiple charged particles. The change of intensity of the primary flux with latitude, therefore, determines the energy spectrum of the particles below these values.

For still higher energies, deflections by the earth's magnetic field become insignificant, and one has to deduce the energy of the primaries from the characteristics of nuclear interactions which they produce. At these energies, collisions frequently give rise to a large number of particles with relativistic energy, which are called shower particles when observed to emanate from the center of the starlike configuration of tracks which indicate that a nuclear interaction has occurred inside an emulsion. It is now known that these shower particles consist mostly of pions (π-mesons). Early attempts to deduce the energy of the primary particle from the number

pions produced in the interaction, or from
ie angular spread of the resulting meson
ıower (the angular divergence between the
ons), were not very successful. It seems that
ultiplicity and angular divergence depend not
ıly on the energy of the incident particle, but
so on other parameters of the collision. The
ost reliable method which is available at
resent for measuring primary energies between
^{10} and 10^{14} ev is based on the observation that
substantial fraction of the collision energy goes
to neutral pions which decay immediately
to γ-rays. These γ-rays initiate an electron-
ıoton cascade which grows rapidly as a result
the well-known alternating process of pair
eation by γ-rays and radiation of γ-quanta by
ectrons. The rate of growth of such cascades
epends on the energy of the γ-ray which
itiates it. It is, therefore, possible to determine
irly reliably, the energy which was given to
ie neutral pion component from the subsequent
owth of an electron-photon cascade. If one
ultiplies this number by three, one has an
timate for the energy given to all (positive,
egative, and neutral) pions and obtains, thereby,
lower limit for the primary energy. It is still
ncertain what fraction of the total primary
ıergy is converted into pions in a high energy
uclear interaction; it seems likely that it
presents between 30 and 50 pct of the incident
ırticle energy.

In spite of this remaining uncertainty, the
ıergy distribution of the various components
the primary cosmic radiation is now fairly
ell known. The experiments described here,
ıd others of a similar nature, have led to
sults of great importance for subsequent
vestigations of the origin of cosmic radiation:

1. The cosmic radiation consists mainly of
uclei of those elements which are most promi-
ent in nature. The composition differs, however,
om the so-called universal abundance of
ements in a significant way, which will be
scussed below in more detail.

2. At comparable energy, primary electrons
onstitute less than 1 pct of the nuclear com-
ponent.

3. Apart from disturbances produced by solar
tivity, the flux of primaries is constant in
me, indicating that the flux is isotropic; that is,
ırticles reach the earth uniformly from all
rections.

4. Primary nuclei belonging to different elements have identical energy spectra. The number of incident particles of atomic number Z with an energy which exceeds ϵ Bev/nucleon decreases with the $\frac{3}{2}$ power of the energy and is given by

$$N_Z\,(>\epsilon) = \frac{K_Z}{(1+\epsilon)^{1.5}}$$

$$\text{for} \quad 0.5 < \epsilon \lesssim 1000 \; Bev/\text{nucleon}.$$

Ultrahigh energy particles—Primaries in the 'ultrahigh energy' region (above 10^{13}–10^{14} ev/nucleon) are so rare that the chance of observing them in emulsions during a balloon flight is quite small. Fortunately, however, the arrival of primaries with such a high energy can be recorded on the surface of the earth because it produces effects which are distinguishable from the effects of lower energy primaries. They give rise to large electron-photon showers in the atmosphere. Since the density of air and the atomic number of its main constituents are low, these cascades grow slowly, penetrate deeply into the atmosphere, and spread over large areas. An extensive air shower (the simultaneous arrival on earth of a large number of electrons) signifies, therefore, the arrival of a very energetic primary particle at the top of the atmosphere.

The large area over which these showers can be detected compensates for their rarity. The number of electrons in the shower furnishes a measure of the primary energy. Unfortunately one cannot, as yet, establish the atomic number of the primaries which produce these events. Experiments to determine the size, the frequency, and the direction of arrival of extensive air showers have been carried out in the Pamir Mountains, in Harwell, in Cambridge (USA), in Moscow, and in Sydney. They consist of sampling the space density of simultaneously arriving electrons over areas of the order of 1 km^2. The experiments are therefore carried out on a gigantic scale; one of the Russian experiments involves sampling of the area by 5000 hodoscoped counters, while in the Cambridge experiment the electron density is sampled by 12 plastic scintillator discs of 1 m^2 area each.

The largest shower detected so far contained about 10^9 simultaneously arriving particles. The energy of the primary particle must have

been close to 10^{18} ev; even if the particle was a complex nucleus its energy could not have been less than 10^{16} ev/nucleon. For comparison, a proton-synchrotron with a magnetic field of 20,000 gauss and a radius as big as that of the earth could accelerate protons to 4×10^{15} ev or heavy nuclei to 2×10^{15} ev/nucleon.

The energy spectrum of the primary particles which produce air showers continues to follow a power law; their number decreases approximately inversely proportional to the square of their energy.

Even at extreme energies, at least up to 10^{17} ev, the primary flux remains independent of sidereal time, indicating that the particles arrive with equal frequency from all directions of space.

As a result of all these measurements on the flux and energy distribution of primary cosmic ray particles, we can derive the following additional conclusions:

5. The energy of primary particles extends from below $\sim 10^{8}$ to at least a few times 10^{17} ev.

6. The flux of particles with kinetic energy greater than 5×10^{8} ev/nucleon is

3100 particles/m^2, sec, steradian,

or 4000 nucleons/m^2, sec, steradian.

7. Their average kinetic energy is 3.6×10^{9} ev/nucleon.

8. The corresponding cosmic ray energy falling on 1 cm^2 of area (near the poles) corresponds to 7×10^{-3} ergs/sec. The energy density of cosmic radiation (excluding rest-mass), in the neighborhood of the earth, is 0.6 ev/cm^3 or $\sim 10^{-12}$ ergs/cm^3. This is approximately equal to the energy density of starlight in our neighborhood.

9. In contrast to starlight, which is strongest in the direction of the galactic plane (Milky Way), the cosmic radiation is completely isotropic in direction.

Before using this information to draw conclusions of an astrophysical nature, it is interesting to make a comparison between the relative abundance of various elements in the cosmic radiation and the abundance of the same elements in the universe as deduced from other observations. The cosmic ray values in Table 1 represent the relative magnitude of the flux of particles with different atomic number but equal energy per nucleon. The data of He, C, N, O, and and $Z > 10$ are correct to about 25 pct. The remaining abundances, as well as the data on universal abundance, are probably correct within a factor 1.5. The latter are compiled mostly from the analysis of stellar spectra and from the chemical composition of the earth's crust and of meteorites.

TABLE 1—*The chemical composition of the cosmic radiation*

Element	Relative abundance (with respect to hydrogen) of the most prominent elements in cosmic radiation	Universal abundance as given by Suess and Urey
Hydrogen	100,000	100,000
Helium	7,000	7,500
Lithium, Beryllium, and Boron	35	3×10^{-3}
Carbon	190	10
Nitrogen	120	15
Oxygen	190	50
$10 \leq Z \leq 30$	140	30
$Z \geq 30$	<0.1	8×10^{-4}

Many more nuclear species can definitely be identified in the cosmic radiation, especially neon, silicon, magnesium and iron. However, even the small amount of atmosphere which the particles must penetrate before reaching balloon-borne equipment is sufficient to modify appreciably the chemical composition, by partial destruction of heavy primaries in nuclear collisions. It is therefore not yet possible to give individual intensities for the less prominent heavier elements.

Compared with hydrogen and helium, the heavy nuclei are five to ten times as abundant in cosmic radiation as on the surface of most stellar bodies. Similar to the universal abundance, cosmic ray abundance falls quickly for elements heavier than those of the iron group.

There is one other important difference between the two abundances. Nuclei of the elements lithium, beryllium, and boron are extremely rare in the universe, but in cosmic radiation they do occur. Although most of those observed at balloon altitude can be accounted for as break-up products of heavier nuclei in the upper atmosphere, there remains a residue which must be attributed to collision processes which occurred before the cosmic rays entered

the atmosphere. These collisions may have occurred either during the acceleration process itself, or during the passage of cosmic radiation through interstellar space. The radiation observed on earth thus bears in its composition traces of its past history.

On the life history of primary particles—It is now possible to draw some conclusions of an astrophysical nature:

1. The presence of nuclei of the iron group, not yet destroyed by nuclear collision, indicates that the particles have not traversed a large amount of matter before they reach the earth. The smallness of the lithium, beryllium, and boron component, arising as a result of collisions and break-up of heavier primaries in interstellar space, shows that the layer of matter, mostly hydrogen, traversed by the primary particles before arrival at the earth weighs less than ~one g/cm^2. This is equivalent to the following statement: The average age of cosmic ray particles arriving in the neighborhood of the earth lies between $10^5/\rho$ and $10^6/\rho$ years, where ρ is the average number of hydrogen atoms/cm^3 along their trajectory.

2. Since the amount of matter traversed by the average particle is small, the chemical composition of the primary particles observed at the top of the atmosphere is nearly the same as that with which they were emitted from the source.

3. Since the ratio of helium (which possesses a very high ionization potential) to elements with much lower ionization potentials is not much smaller than that corresponding to their relative natural abundances, the acceleration must, at least initially, take place in a hot and rarefied region, a region where the majority of atoms, even most helium atoms, are ionized. When combined with conclusion (1), namely, that the gas density in the source region must be low, it follows that the source is most probably to be found near some stellar bodies, yet at some distance from their surface.

4. The isotropic distribution in the directions of arrival of primary particles suggests that at least the particles with very high energy cannot have originated in the sun and therefore probably not in the other ordinary stars of the main sequence. (The contribution which the sun makes to cosmic ray intensity is discussed in Section IV.) We must look for possible sources in exceptional stars. Only if we fail to find such stars will we be forced to look for the source of cosmic radiation outside of our galaxy.

5. Since stars in our galaxy are strongly concentrated in the galactic plane, and further concentrated within the plane towards the galactic center, the distribution of sources is not likely to be isotropic with respect to the earth. If, nevertheless, the particles arrive with equal frequency from all directions, they must have been scattered through large angles in interstellar space. The deflections must be due to extended interstellar magnetic fields. In order to be effective in randomizing the direction of even those primary particles whose energy is high, the field strength cannot be much smaller than 10^{-6} gauss.

There are no plausible alternative mechanisms to account for the isotropy of incident primary particles. Scattering by the mechanism of nuclear collision is ruled out by the presence of complex nuclei among the primaries. Extended strong electric fields cannot be responsible; they cannot build up because of the high electric conductivity of interstellar space. Interstellar magnetic fields were first predicted by Fermi on the basis of cosmic ray evidence. Their existence has since been verified through other phenomena: the partial polarization of starlight by the interstellar medium, and the emission of radio noise from the galaxy.

III. Cosmic radiation in our galaxy

Regions containing cosmic rays—At this point, the information acquired through the study of primary cosmic rays and that accumulated by the rapidly developing new science of radio astronomy began to converge towards a picture, in which both, at first sight quite unrelated, phenomena are closely linked. The study of the isophotes of galactic radio noise revealed that the galaxy is surrounded by a roughly spherical region filled with a large number of diffuse, weakly emitting centers which cannot be identified with visible astronomical objects. It became apparent that not 'radio stars,' but concentrations of electrons spiralling around magnetic lines of force must be responsible for radio emission. It was discovered also that other galaxies, in particular M31, a flat spiral galaxy similar to our own, are surrounded by spherical halo regions which emit radio noise. As a result of these observations it is now generally accepted

that our own galaxy and other similar galactic systems are surrounded by halos of turbulent, magnetized gas clouds. In our galaxy the halo has a radius of about 50,000 light years. It seems reasonable to assume that this is the volume in which the cosmic ray particles exist and are prevented from escaping by deflection in irregular magnetic fields.

In a highly conducting medium, such as the interstellar gas, the decay of magnetic fields is so slow that it becomes negligible even on an astronomical time scale, and one may consider the fields as frozen into the clouds, partaking in their motion. This fact can be used to estimate the strength of these fields in relation to the density, ρ, and velocity, v, of interstellar gas clouds. Magnetic fields in clouds of turbulent gas cannot, for a long period, exceed a value given by

$$H^2/8\pi = \tfrac{1}{2}\rho v^2;$$

otherwise the repulsion between magnetic lines of force would react on the motion of the clouds so as to reduce the turbulent velocity. On the other hand, in a turbulent gas of high electrical conductivity, magnetic lines of force become entangled and drawn out so that their density increases even if the initial density was low. Thus, we may assume that an equilibrium has been established between the energy density of the magnetic field, $(H^2/8\pi)$, and the kinetic energy, density $(\frac{1}{2}\rho v^2)$, of the clouds.

In order to maintain the spherical halo it is necessary that, at any radius, the pressure exerted by the magnetic field and the pressure exerted by the cosmic radiation which tries to escape through this field is balanced by the weight of the layer lying above it. Since the gravitational potential of the galaxy is known, it is possible to make at least a rough estimate of the gas density and field strength in the halo. The values set down here are those estimated by Shklovsky:

At a distance r from the galactic center

$H = 3 \times 10^{-6}$ gauss, $n = 6 \times 10^{-3}$ atoms/cm^3

for

$$r > 30{,}000 \text{ light years}$$

$H = 6 \times 10^{-6}$ gauss, $n = \sim 10^{-2}$ atoms/cm^3

for

$$r < 30{,}000 \text{ light years.}$$

In the optically observable disc-shaped regio of the galaxy one finds, in addition to chaoti fields, orderly fields of about 8×10^{-6} gaus which extend along the spiral arms. In this dis which comprises roughly 1 pct of the galact volume, the mean gas density is about 1 atom cm^3.

The equipartition between magnetic fiel energy and the kinetic energy of turbulenc requires then that the hydrogen clouds in th halo move with velocities of the order of 10 km/sec. Velocities of this magnitude are als indicated by the size of the halo; in order t move out a distance of the order of 50,000 ligh years against gravitational attraction, a clou starting from the galactic plane must have velocity of about 150 km/sec.

The origin of galactic radio noise—If cosmic ra particles move through galactic space, they wi occasionally collide with interstellar hydroge If their energy is high enough, a substantial par of the energy of colliding nucleons (at least 3 pct and perhaps more) will be converted int high energy pions. One third of these pions ar neutral; they decay immediately into two γ-ray and their energy will be lost to our galaxy. Th remaining two thirds (at least 20 pct of th energy of the colliding nucleons) are carried b charged pions, which decay via μ-mesons int electrons and neutrinos according to the schem

$$\pi^{\pm} \rightarrow \mu^{\pm} + \nu$$

$$\mu^{\pm} \rightarrow e^{\pm} + 2\nu$$

On the average, the electrons will receive 26 pc of the pion energy. Therefore, high energ electrons and positrons will be continuousl produced wherever cosmic radiation exists, an these electrons will have an energy spectru similar to that of the primary nuclei whic produce them. They will obtain at least $0.2 \times 0.26 = 5$ pct of the energy of the collidin nucleons. If magnetic fields are present, th electrons will spiral around the lines of force an lose energy primarily by radiation. Is it possibl that these electrons which arise from the inte actions of primary nuclei in the interstellar ga are responsible for the galactic radio noise There are two arguments which seem to favo such an explanation.

If the magnetic field strength in the galacti halo lies between 10^{-5} and 10^{-6} gauss, the radia

on of radio noise requires electrons of about
ne Bev energy.

The total power radiated by the galaxy in
dio wavelengths can be estimated (within a
ctor of two) to be

$$P = 2.5 \times 10^{38} \text{ ergs/sec.}$$

he energy transfer from cosmic ray nuclei to
ie electron component has very nearly the
me value if we assume that the average cosmic
y density in the halo is that observed in our
eighborhood and that the mean gas density is
$\sim 10^{-2}$ hydrogen atoms/cm^3. (We use here our
revious minimum estimate that 5 pct of the
llision energy will appear in the electron
mponent.) In order to produce this power,
ie density of electrons at a given energy should
e about 100 times smaller than the correspond-
g density of cosmic ray protons. This is
nsistent with the previously quoted observation
iat the electron flux in the primary radiation is
)t larger than 1 pct of the proton flux.

The second argument in favor of the hypothesis
iat cosmic ray nuclei may be responsible for
ie radio emission of the galaxy comes from the
)ectral distribution of the radio noise. The
)wer radiated from the galactic halo decreases
ith increasing frequency, ν, as $\nu^{-0.8}$. To produce
ich a spectrum requires an energy distribution
electrons of the form $n_e(> E) = K_e/E^{1.6}$,
hich is very similar to that of cosmic ray nuclei.
his is suggestive, but is not a decisive argument.
)r, even if one assumes that the electron
)ectrum at production is identical with that of
ie cosmic ray nuclei, the competing effects of
iergy losses by radiation, energy gains by
llision with inhomogeneities in the magnetic
eld, and the possible escape of electrons from
ie galaxy will tend to modify their energy
)ectrum with time. Thus, the evidence seems
irly strong, but not conclusive, that part or all
the radio emission in the galactic halo is
directly the result of nuclear collisions of
)smic ray nuclei in the interstellar gas.

Sources of cosmic ray particles—It now seems
asonable to look for a source of cosmic ray
imaries in those regions of the sky where the
ost intense radio sources are located. One of the
rongest radio sources in our galaxy coincides
ith the luminous gas cloud known as the Crab
ebula. It is a rapidly expanding cloud, the
sult of a super nova explosion of the year
1054 A.D. The light emitted by the cloud has long been a puzzle to astronomers because of the absence of strong spectral lines. Is it possible that not only the radio emission but also the light emitted in the optical region are due to the same cause, high energy electrons spiralling in magnetic fields?

This can be tested experimentally, because, in contrast to the radiation from excited atoms and molecules, the light emitted by electrons spiralling in magnetic fields is completely polarized. The discovery that the light emitted from the Crab nebula is indeed strongly polarized, represents strong evidence that we are dealing here with synchrotron radiation of relativistic electrons. One must now accept as certain that there exist regions in our galaxy where particles in large numbers are accelerated to relativistic energies. Because it is difficult to think of a mechanism which will accelerate electrons but not ions, it seems reasonable to assume that cosmic ray nuclei are also accelerated and receive an energy comparable to, or larger than, that given to the electron component. The nuclei are, however, not directly observable because, owing to a larger rest-mass, their electromagnetic radiation is negligible.

It is now of great importance to estimate the number and energy of electrons which give rise to the strong radio emission and optical synchrotron radiation in super novae, in order to see whether the acceleration of a comparable number of nuclei would represent an adequate source for the cosmic radiation which is present in the galaxy.

The spectrum and power radiated by an electron spiralling in a magnetic field are known from electromagnetic theory. Our distance from the Crab nebula is known to be $\sim$4000 light years and the radiated power is

$F_\nu = 1.5 \times 10^{-23}$ watts/cm^2, steradian, (c/s), in the optical region ($\lambda = 4300$ A)

and

$F_\nu = 1.8 \times 10^{-20}$ watts/cm^2, steradian, (c/s), in the radio band at 100 Mc/s.

A field of $\sim 10^{-4}$ gauss perpendicular to the motion of the particles corresponds then to the following values for the number, N_e, and energy, W_e, of electrons: The electrons responsible for the optical radiation at $\lambda = 4300$ A have energies $E \sim 10^{12}$ ev, their number is $N_e \approx 10^{47}$, and their energy $W_e \approx 10^{47}$ ergs. The electrons

responsible for the emission of radio noise at 100 megacycles have energies $E \approx 5 \times 10^8$ ev, their number is $N_e \approx 10^{50}$, and their energy $W_e \approx 10^{47}$ ergs. These are minimum estimates based on the assumption that the energy of the electrons is just right to give the maximum contribution to the radiation at these wavelengths. The true values could be higher by a factor of ten.

These estimates are based on a magnetic field strength of 10^{-4} gauss. The field could be smaller, but then the number and energy of electrons responsible for the observed radiation would have to be larger than estimated above. The field could also be larger than 10^{-4} gauss, which would reduce the number and energy of electrons required for the emission of radiation; but in the stronger field, the electrons, instead of losing 50 pct of their energy by radiation in 1000 years, would lose their energy in a much shorter period. The total energy transformed by the super nova into relativistic particles would, therefore, not be reduced; and, instead of assuming that the acceleration of particles dates back to the original explosion, we would then have to assume that it is still taking place today.

There are other radio sources in our galaxy. The strongest, in Cassiopeia, is believed to be the result of a super nova outburst of 369 A.D. On the basis of arguments similar to those given above, one estimates that it contains about 10^{52} electrons of energy above 5×10^8 ev, about 100 times more than the Crab nebula.

If one assumes that cosmic ray nuclei are produced in comparable numbers with the electrons, one must interpret these and similar sources of radio noise and of polarized light as representing a number of presently active accelerators of cosmic ray particles. Is their power output large enough to account for the observed cosmic ray intensity?

If the density of cosmic radiation in the galactic halo were the same as that observed in our neighborhood, the energy stored in the galaxy as cosmic ray nuclei would be

$$W_{CR} \approx 10^{-12} V \approx 10^{56} \text{ ergs},$$

where $V \approx 10^{68}$ cm^3 is the estimated volume of this halo. The average time for cosmic ray protons to lose an appreciable fraction of their energy as a result of a nuclear collision depends on their collision cross section, σ, and on the mean density, n, of interstellar hydrogen gas. Using $\sigma = 35$ millibars and $n = 10^{-2}$ hydrogen atoms/cm^3, this time is

$$\tau = \frac{1}{n\sigma c} \approx 3 \times 10^9 \text{ years},$$

which is somewhat smaller but comparable to the age of our galaxy. Therefore, if the particles do not leak out into extragalactic space, it requires $W_{CR}/W_e = 10^6$–10^7 outbursts of the Cassiopeia type (on the average one outburst every 300 to 3000 years) to produce sufficient cosmic ray energy. If the average energy of a super nova outburst is only of the order of 10^{47} to 10^{48} ergs, as in the Crab nebula, it would require one such explosion every 3 to 30 years.

The frequency of super nova outbursts in our galaxy is not well known. During the last 1000 years, five such outbursts have occurred within ~7000 light years from the solar system. This would indicate a frequency in the entire galaxy of somewhat more than one outburst every 100 years. Super nova explosions may, therefore, provide an adequate source of cosmic ray energy, provided the particles do not escape easily through the galactic boundary.

The assumption that cosmic ray nuclei are kept in the galaxy indefinitely until they make nuclear collisions seems to contradict our previous conclusions that the radiation observed on the earth is fairly new and has suffered only very few nuclear collisions since its acceleration. One must, however, remember that both the super novae and the solar system lie in the galactic disc with its spiral arms, where the structure of the magnetic field is quite different from that in the outer halo. On the basis of such considerations it seems possible to devise a reasonable mechanism for the diffusion of particles in the galactic volume which will reconcile these requirements, but a discussion of this problem would lead us too far afield.

There may of course be other important sources of cosmic ray primaries which have not yet been discovered. All one can say at present is that it is practically certain that some cosmic ray particles originate in super novae explosions; it is *possible* that these explosions furnish the major fraction of cosmic ray nuclei.

The mechanism by which cosmic ray nuclei

nd electrons are accelerated in super novae xplosions is not yet well understood. Other alaxies exist with an output of radio power and f polarized optical radiation which indicates hat they contain as much as 1000 times more osmic ray particles and cosmic ray energy than re contained in our own galaxy.

Possible relation between cosmic radiation and volution—It is well known that ionizing radiation nduces mutations. Geneticists seem to agree hat, at its present intensity, cosmic radiation lays only a minor role compared with other nutagenic agents. However, if the solar system ver passes through the expanding gas cloud of super nova explosion, then, as first pointed out y Krassovsky and others cosmic ray intensity nust rise by a large factor, and during such a eriod it would probably influence the mutation ate of long-lived species significantly. The history f the galactic radio sources which are now bservable indicates that, around the center of he explosion, a density of cosmic radiation ten o a hundred times as high as in the surrounding pace persists for a period of perhaps 2000 to 000 years, at which time the expanding gas loud will occupy a sphere with a diameter of 5 to 20 light years. The probability of a super ova explosion occurring so close to the solar ystem is not negligible; it may occur at the ate of once every 200 to 300 million years. If ne therefore accepts the super nova theory of he origin of cosmic rays, it becomes very ikely that there were periods in the earth's istory, lasting for perhaps a few thousand years, luring which cosmic radiation was many times arger than at present. The history of evolution, nd perhaps the history of life itself, may there- ore be closely connected with the past intensity f cosmic radiation.

IV. Cosmic radiation in the planetary system

High energy particle emission from the sun—On ebruary 23, 1956, at 3 hours 44 ± 1 min. GMT, osmic ray recording equipment in Europe and sia registered an enormous increase in the umber of particles incident on the earth's urface. Counting rates rose quickly to values p to 50 times higher than normal. The increase ecorded by each station depended in a systematic anner on its latitude and altitude, as well as on the nature and energy of the particles which were detected.

About ten minutes later similar increases occurred in North America and Canada. Peak values were reached within a few minutes and were followed by a slow return to normal conditions attained nearly fifteen hours later.

This occurrence was evidently connected with a large solar flare of magnitude 3+ recorded in Japan and in India. The flare had started at 3h 31m GMT and died out 45 minutes later. It was accompanied by a radio fadeout.

Although this was the largest and most completely documented increase in cosmic ray intensity ever recorded, it was not the first one. Altogether four similar events had been observed in the preceding 14 years, the last one having occurred on November 19, 1949.

A detailed analysis of the records obtained by a network of 40 stations covering many latitudes and longitudes led to the following conclusions:

a. The particles responsible for the earliest increase were not isotropically distributed around the earth. They came from a direction pointing to an area on the celestial sphere which was centered around the sun but was much larger than the solar disc. This can be deduced from the extent in longitude of the initial impact zone, which stretched from Amsterdam in Holland over all of Europe and Asia to Cape Schmidt on the Pacific. Yet this zone was rather well defined and definitely did not include the northern part of the Western Hemisphere.

b. The solar burst contained a much larger percentage of low energy particles than does ordinary cosmic radiation. High energy particles, however, were also present and the spectrum extended to at least 40–50 Bev. This information was obtained by studying the amplitude of the increase as a function of geomagnetic latitude, and therefore as a function of the primary cut-off energy imposed by the earth's field.

c. Although the velocity of the primary particles, whose secondaries were recorded, must have been very close to that of light, the first particles arrived several minutes later than the light of the flare and the x-ray burst responsible for the radio fadeout. There was, furthermore, an indication that the more energetic particles arrived earlier than those of lower energy, and that the difference in arrival time was much too large to be attributed to a differ-

ence in particle velocity. This information came from comparing the onset time of the burst at stations located in different latitudes.

d. The fraction of energy, liberated in the flare, which was used up in the acceleration of cosmic ray particles was comparable to the energy given off in the form of visible light ($\sim 10^{32}$ ergs).

It seems safe to generalize and extrapolate these conclusions somewhat and restate them in the following form:

1. Explosions on the surface of the sun (and presumably other stars of the main sequence) can give rise to cosmic ray particles with energies up to at least 40 Bev.

2. The accelerating process can be extremely efficient so that a very substantial fraction of the available energy can be converted into cosmic radiation.

3. Acceleration to very high energies can take place in a matter of minutes.

4. Bursts of cosmic ray particles from the sun are rare. If all the stars in our galaxy contributed cosmic ray particles at the same rate as they are produced in solar flares, the combined production could account for only a very small fraction of the cosmic ray intensity which is normally observed.

5. The paths by which the earliest particles travelled were at least twice as long as the direct distance from the sun to the earth. The region between the earth and the sun must therefore have contained magnetic fields strong enough to curl up the trajectories.

6. Since particles continued to arrive many hours after the flare had subsided and were at that time distributed completely uniformly around the earth, magnetic fields capable of scattering these particles must have existed also far beyond the orbit of the earth.

There are other conclusions which one can draw about the strength and distribution of these interplanetary fields, but they are less certain and need support from further observations. Yet it is evident that cosmic ray particles produced in solar flares can give interesting information on magnetic fields in the space within and beyond the earth's orbit.

Intensity modulations originating in interplanetary space—The method for studying interplanetary magnetism by means of large solar flares has serious limitations; suitable events are extremely rare, and when they d occur, the observations refer to highly disturbe rather than to 'normal,' conditions of th interplanetary medium.

Another powerful method, which is free fro these particular disadvantages, is based o modulations (produced by changes in inte planetary conditions) in the flux of norm cosmic radiation, that is, those particles whic arrive from outside the solar system. The stud of such effects involves continuous monitorin of the primary cosmic ray intensity (by means o some secondary cosmic ray effect) at man points of the earth's surface and correlating th observed intensity changes. The largest effec must be expected for primaries of low energy They are not easy to monitor because the earth' field restricts their arrival to the polar caps an because their secondary effects do not exten deep into the atmosphere. Yet, in numerou balloon flights with ion chambers and counter Neher and his collaborators were able to establis that the low energy primary radiation suffe large intensity changes; in particular, he foun decreases by a factor of more than two betwee years of low sunspot activity and years of maxi mum sunspot number.

For long term synoptic observations, one is a yet restricted to monitoring devices on th ground, where only primaries whose energ exceeds several Bev can produce effects an where the amplitude of the modulations i therefore much smaller. In such measurements intensity variations, due to changes outside th atmosphere, are superimposed on changes orig inating within the atmosphere, and the meteoro logical effects are often dominant. The resultin difficulties in distinguishing interplanetary fro atmospheric disturbances have been overcom only during the last few years.

A particularly useful monitoring device is th neutron pile developed by Simpson and hi collaborators. One measures neutrons, locall produced by the nuclear component of th secondary cosmic radiation. As shown in Sectio V, such neutrons are coupled to the primar radiation through a chain of nuclear interaction which ordinarily do not involve the creation o the decay of unstable particles. The minimun energy, which a primary particle must posses so that it can still contribute to the neutro flux, does not, therefore, exceed a few Bev. In thi

nergy range, intensity variations of interplane-
ary origin are still comparatively easy to detect.
urthermore, since no unstable particles are
nvolved, neutron production near the ground
oes not depend on the distribution of air density
ithin the atmosphere and is therefore inde-
endent of temperature effects. The only
eteorological effect which must be corrected for
s that due to the easily measurable changes in
arometric pressure.

If, in place of neutrons, one uses the flux of
-mesons or the total ionization near the ground,
s a monitor of primary cosmic ray intensity,
ne measures intensity changes for primary
articles of higher energy. In that case one has to
ontend not only with the smaller amplitude
f the variations, but also with more complicated
eteorological effects which are difficult to
liminate from the data.

Continuous monitoring of cosmic ray intensity
uring the last years has led to the conclusion
hat the activity on the surface of the sun governs
he transparency of interplanetary space for
osmic ray particles and that this transparency
an be quite low for particles with one Bev
nergy or less. The more active the sun, the
ower the transparency and the smaller the
ux in the low energy part of the spectrum. This
s most clearly shown by the following observa-
ions:

1. Solar flares which produce radio fadeout and
agnetic storms may lead to a sudden world-
ide drop of cosmic ray intensity, which then
ersists for several days (Forbush decreases).
he magnetic storms are not, however, the
ause of the decrease. Storms can occur without
drop in cosmic ray intensity; also the latitude
ependence of decreases is such that they could
ot be produced by simply changing the geo-
agnetic cut-off energy of the incident radiation.]

2. Certain cosmic ray intensity variations
ave a tendency to recur at 27 day intervals;
hey may return up to nine, or even eleven
imes, before they die out. Several such quasi-
eriodic disturbances can exist concurrently.
wenty-seven days is the apparent rotational
eriod of the sun as viewed from the earth.)

3. The general level of cosmic ray intensity is
ower in years with a high sunspot number than
n the years when the sun's surface is relatively
uiet. The change is large (a factor of two or
nore) for the low energy primaries in the polar stratosphere. It is smaller for particles of higher energy; for those primaries whose energy is high enough to contribute to the meson flux at sea level (~20 Bev), it amounts to approximately 4 pct.

From these observations one can conclude that the ionized gas clouds, which are expelled from the solar surface during an eruption, carry magnetic fields capable of reflecting incident cosmic ray particles. If the interplanetary space is filled with such clouds, low energy particles can reach the earth only after a kind of diffusion through the interplanetary medium. To be effective, the magnetic fields must be of the order of 10^{-4} gauss or higher. Presumably the clouds are identical with the plasma clouds which may encompass the earth and then produce magnetic storms. They usually arrive 20 to 24 hours after the corresponding radio fadeout, which shows that they move away from the sun with a speed of about 2000 km/sec.

In order to account for the long term change of cosmic radiation intensity with the 11-year period of sunspot number, one must assume that the effect of these clouds on cosmic ray particles is cumulative; the lifetime of an individual magnetized plasma must therefore be considerable, and clouds must exist in regions far outside the earth's orbit.

Evidently, the information on the distribution of fields and matter in the interplanetary space, which one has obtained so far by monitoring cosmic ray intensity, is only qualitative and lacks precision. But there seems ground for hope that more detailed investigations of this kind will ultimately lead to a fairly accurate knowledge of interplanetary conditions.

Cosmic ray evidence for the influence of interplanetary plasmas on the earth's magnetic field.— Apart from the well-known fact that highly disturbed conditions in the interplanetary plasma can affect the earth's field by producing magnetic storms, cosmic ray investigations have given an indication that, even in the absence of such disturbances, the earth's magnetic field may be distorted by interaction with the highly conducting interplanetary gas. The geomagnetic field, a few thousand kilometers above the surface of the earth, differs from the field which one calculates by using the measured magnitude and direction of field vectors on the ground. This is most clearly demonstrated by the fact

that the line of minimum cosmic radiation intensity on the surface of the earth does not coincide with what was believed to be the geomagnetic equator. This discrepancy seems to indicate that there is something wrong with the geomagnetic calculations. They have been carried out on the assumption that the earth rotates in a vacuum; the discrepancy may be an indication that this approximation is not as good as had previously been assumed.

V. Cosmic radiation in the atmosphere

The primary nuclei beat against the top of the atmosphere like heavy rain; about one particle falls on a square centimeter every second. The earth's surface is protected from the bombardment by a layer of air, equivalent in its protective power to 10 meters of water. The processes by which the enormous energy of individual nuclei is transformed and dissipated in this layer are quite complicated and their study in the last decade has given rise to a new branch of physics: high energy or particle physics.

Modern detectors have made it possible to distinguish a great variety of particles in the air and to establish the principal generic relations between them. As a result of this work one can account now, with reasonable accuracy, for the energy distribution and intensity of the various secondary components of the radiation in different layers of the atmosphere. The most important single step which has led to an unravelling of these relations was the detailed study of collisions between high energy nuclei, a study which was made possible by improved instrumentation. The instruments which proved best suited for identifying the large variety of new particles which are created in high energy collisions are those in which the collision and decay events themselves can be made visible: cloud chambers and nuclear photographic emulsions. The former are more useful for studies near the ground, the latter for exploring the upper regions of the atmosphere.

The particles produced in the atmosphere as a result of cosmic ray bombardment represent a complete catalogue of all the particles known to physics. Their names and their weight compared to that of the electron are listed in Table 2. If they are unstable, their lifetime and principal modes of decay are given. Particles whose existence was first established in cosmic radiation are marked with asterisks. Of the 21 particles listed in Table 2, twelve were discovered in the cosmic radiation, and strong evidence for the existence of three of the others (π^0, $\bar{p}$, $K_2{}^0$) was obtained before their existence was finally proved with the help of particle accelerators.

In addition to the particles listed here, there exist in the atmosphere the products of transformation of those air nuclei which have been involved either in a high energy collision or in reactions with secondary cosmic ray particles. They include the various stable and radioactive nuclei which can be produced in nitrogen, oxygen, carbon, argon, neon, xenon, and krypton, by partial destruction, by neutron capture, or by the capture of negative μ-mesons. They also include such oddities as the so-called hyperfragments, nuclei which are radioactive by virtue of the fact that one of their constituent nucleons is replaced by an (unstable) lambda hyperon.

Although all these particles occur in the atmosphere, most of them are quite rare and play only a minor role in the particle cascades initiated by the primary nuclei. Their importance lies in the fact that their discovery has led to a great surge of interest in the field of particle physics, that it has stimulated the invention and construction of ever more powerful particle accelerators, and that this increased activity promises to lead us to an understanding of the nature of matter much deeper than we have at present.

Phenomenologically the most important particles in the cosmic radiation, apart from primaries, are the secondary nucleons (protons and neutrons) and the tertiary component consisting of μ-mesons, γ-rays, and electrons. The pions form the main link between the nucleons and the light particles, and their discovery, by C. F. Powell and his colleagues at Bristol, proved to be the key to the understanding of secondary cosmic ray phenomena. The lifetime of pions is so short that they never accumulate in substantial numbers in any part of the atmosphere. Roughly speaking, nucleons predominate in the uppermost 5 pct of the atmosphere; electrons, positrons, and γ-rays predominate in the pressure region between 100 and 600 g/cm^2 (between 5 and 17 km), and μ-mesons in the lower region, as well as below the surface of the earth. This distribution is a consequence of the properties of the particles produced in high energy

collisions. The interpretation of cosmic ray observations in the atmosphere can be summarized as follows:

Collisions between nuclei lead to fragmentation, which means the emission of neutrons, protons, and more complex fragments. When the energy of the incident particle exceeds ~1 Bev/nucleon, pion production sets in. Most of the energy lost by the incident particle is transformed into pions, of which two thirds belong to the charged, and one third to the neutral variety.

The charged pions decay into μ-mesons according to the scheme shown in Table 2 and the resulting μ-mesons, if not energetic, will be slowed down by ionization loss in the air and decay into electrons or positrons. However, if the μ-meson energy exceeds about 2 Bev, most of them will reach the earth.

Three reasons cooperate here. First of all, the μ-meson now possesses more energy than it will lose in the atmosphere by ionization. Secondly, its velocity is so high that the relativistic effect of time dilation becomes important; this leads to an increase in its mean life by a factor of twenty or more, and permits it to cover the distance from its place of origin to the ground before decay. Thirdly, μ-mesons, in contrast to nucleons and pions, interact only very weakly with other particles. A particle which traverses a layer of air corresponding to more than 60 g/cm^2 is likely to run into a nucleus of nitrogen or oxygen. Fast protons, neutrons, and pions lose a large fraction of their energy in such encounters. Nucleons of a few Bev have, therefore, little chance of traversing the atmosphere (~1000 g/cm^2). On the other hand, μ-mesons can pass freely through nuclear matter, in spite of its high density of 3×10^{14} g/cm^3. The effects which they produce in the air are those due to their electric charge, which leads only to disturbances in the electronic arrangement of the atoms which they pass. Loss of energy due to this weak electric interaction is comparatively small. As a result of their weak interaction with matter, μ-mesons predominate in cosmic radiation at great atmospheric depth.

Returning to the primary collisions of nuclei, we see from Table 2 that those pions which are electrically neutral decay into γ-rays and not μ-mesons; the fraction of cosmic ray energy which enters into the neutral pion component has, therefore, an entirely different fate. High energy γ-rays produce electron-positron pairs when they pass close to a nucleus. On the average, ~50 g/cm^2 of air are needed to effect such a conversion. The resulting electrons and positrons radiate quanta when they pass close enough to a nucleus to suffer a large deflection in its coulomb field. These γ-rays in turn will produce more electron-positron pairs, etc. Thus, the energy given to neutral pions leads to an electron-photon cascade. The original energy is rapidly distributed among many particles, and since each particle loses energy by ionization, the total loss by ionization increases rapidly; finally the cascade ceases its growth and dies out for lack of energy. The rapid multiplication of particle number in this process is responsible for the strong maximum in the secondary cosmic radiation at an atmospheric depth of about 150 g/cm^2 (or 14 km), where electrons represent the strongest component.

Compared with these processes, which are connected with the production and the decay of charged and neutral pions, other nuclear and electromagnetic processes play a minor role in cosmic radiation. Secondary protons and neutrons emitted in the primary acts of nuclear disintegration produce further nuclear reactions of smaller energy. A nucleon cascade results, in which protons and neutrons play identical roles, until the energy falls below about 500 Mev. Thereafter, most protons are brought to rest by ionization, while the neutrons continue to produce smaller and smaller nuclear reactions. Finally, when they approach thermal energy, the neutrons are captured by N^{14} to produce the radio isotope C^{14}.

If a proton at the top of the atmosphere has an energy in excess of 10^{12} ev, it will reach the surface of the earth with an energy of a few Bev, in spite of undergoing several pion producing collisions. New electron-photon cascades and nucleon cascades will start at each of the points in the atmosphere where the primary particle suffers a collision. Thus, by the time it reaches the ground, the primary will be accompanied by a number of electrons, μ-mesons, and nucleons. If the energy of the primary is very high, the number of shower particles will be large. Coulomb scattering and angular divergence at production spread the electrons and mesons over a large area and lead to the phenomenon which is called an extensive air shower and which has already been mentioned in the earlier section.

Table 2

Symbol	Name	Rest mass (electron masses)	Mean life (seconds)	Decay products
γ	Photon	0	Stable	
ν	Neutrino	0	Stable	
e^-	Electron	1	Stable	
* e^+	Positron	1	Stable	
* μ^-	μ-meson	206.9	2.2×10^{-6}	$\mu^- \rightarrow e^- + 2\nu$
* μ^+	μ-meson	206.9	2.2×10^{-6}	$\mu^+ \rightarrow e^+ + 2\nu$
* π^-	Pion	273.3	2.6×10^{-8}	$\pi^- \rightarrow \mu^- + \nu$
* π^+	Pion	273.3	2.6×10^{-8}	$\pi^+ \rightarrow \mu^+ + \nu$
π^0	Pion	264.4	10^{-16} to 10^{-15}	$\pi^0 \rightarrow \gamma + \gamma$
* K^-	K-meson	966.5	1.2×10^{-8}	$K^- \rightarrow \mu^- + \nu$, $\pi^- + \pi^0$, $\pi^+ + \pi^- + \pi^-$,
* K^+	K-meson	966.5	1.2×10^{-8}	$\pi^- + \pi^0 + \pi^0$, $\pi^0 + \mu^- + \nu$, $\pi^0 + e^- + \nu$
* K_1^0	K-meson	965	1×10^{-10}	$K_1^0 \rightarrow \pi^+ + \pi^-$, $\pi^0 + \pi^0$
K_2^0	K-meson	965	4×10^{-8} to 1.3×10^{-7}	$K_2 \rightarrow \pi^- + e^+ + \nu$, $\pi^- + \mu^+ + \nu$; ${}^0\pi^+ + \pi^- + \pi^0$, $\pi^0 + \pi^0 + \pi^0$
n	Neutron	1838.6	1000	$n \rightarrow p + e^- + \nu$
$\bar{n}$	Anti-neutron	1838.6	(1000)	$(\bar{n} \rightarrow \bar{p} + e^+ + \nu)$
p	Proton	1836.1	Stable	
$\bar{p}$ (p^-)	Anti-proton	1836.1	Stable	
* Λ^0	Lambda Hyperon	2182	2.8×10^{-10}	$\Lambda^0 \rightarrow p + \pi^-$, $n + \pi^0$
* Σ^-	Sigma Hyperon	2341	1.7×10^{-10}	$\Sigma^- \rightarrow n + \pi^-$
* Σ^+	Sigma Hyperon	2328	8×10^{-11}	$\Sigma^+ \rightarrow p + \pi^0$, $n + \pi^+$
Σ^0	Sigma Hyperon	2337	$< 10^{-11}$	$\Sigma^0 \rightarrow \Lambda^0 + \gamma$
* Ξ^-	Xi-Hyperon	2585	4.6×10^{-10} to 2×10^{-8}	$\Xi^- \rightarrow \Lambda^0 + \pi^-$

The production and decay of pions, the properties of μ-mesons, the processes of pair creation by γ-rays, and of bremsstrahlung by electrons form the key to the understanding of cosmic ray phenomena in the atmosphere. It seems that all existing observations on cosmic radiation in the atmosphere can be understood reasonably well, both qualitatively and quantitatively, in terms of these processes and the particles which we have discussed.

VI. Cosmic ray produced radioactivity

Within the last decade, another aspect of cosmic radiation studies has gained in importance. It is the study of the radioactivity which is produced in the atmosphere in the wake of nuclear disintegrations. This radioactivity is very weak compared with that of most other natural or artificial sources; it is nevertheless useful, because it provides a steady source of tracer elements whose half-life is short compared with that of the naturally radioactive substances which exist on the earth.

These tracer elements can be used for studying the exchange of material between various layers of the atmosphere or of the oceans, and also exchanges between atmosphere, biosphere, and hydrosphere. A list of the activities which have been identified so far is given in Table 3. The half-life of each isotope and the principal modes of production are also listed. These isotopes cover a wide range of half-lives. Others will no doubt be added to this list.

The production rate of radio isotopes varies with the number of nucleons in the cosmic radiation and is therefore a function of altitude,

TABLE 3

Isotope	Half-life	Mode of production
Be^{10}	2.7 million years	Spallation of N^{14}, O^{16}
C^{14}	5600 years	Neutron capture in N^{14}
H^3	12.5 years	Spallation of N^{14}, O^{16}
Na^{22}	2.6 years	Spallation of A^{40}
S^{35}	87 days	Spallation of A^{40}
Be^7	53 days	Spallation of N^{14}, O^{16}
P^{33}	25 days	Spallation of A^{40}
P^{32}	14 days	Spallation of A^{40}
Cl^{39}	1 hour	Spallation and μ-meson capture in A^{40}

as well as of geomagnetic latitude; but, apart from small and short-lived disturbances, it is independent of time.

The global average rate of production varies from 1.8 nuclei per second per kilogram of air, in the case of C^{14}, to about 4×10^{-4} nuclei per second per kilogram of air, in the case of P^{32}. This corresponds to a total yearly global production of 6.8 kg of C^{14} and 3.6 grams of P^{32}. With the exception of C^{14}, the isotopes listed in Table 3 reach the earth, mainly as constituents of rain water. Enough can be extracted from a few litres to make identification possible, in spite of the fact that the concentrations are extremely low. For instance, an average rain contains 40 atoms/cc of P^{32}, or about one in every 10^{21} atoms. This corresponds to a decay rate of one disintegration per minute, due to P^{32}, in every litre of rain water. Yet 0.5 litre is sufficient to identify the isotope.

The technical development which has made it possible to detect such low level radioactivity consists essentially in improvements concerned with keeping unwanted background radiation away from the radiation detector. Such background activity is mainly due to: (a) cosmic radiation, (b) γ-rays from the natural radioactive elements in the earth, (c) surface contaminations of counter walls, and (d) air borne radioactivity. The reduction of background is achieved by: (a) surrounding the sample counter completely with guard counters which eliminate registration of those events which are associated with a charged particle entering from the outside, (b) massive shields of iron and mercury to absorb γ-rays, (c) careful selection of materials used for construction, and (d) making the counters very small and counter walls very thin. In this manner, the smallest amount of radioactivity, which is still measurable, has gradually been reduced to a value of about one disintegration per hour. Cosmic ray produced radioactivity, though weak, has therefore become detectable.

The first of the cosmic ray produced isotopes to be discovered was C^{14}. Unlike the other isotopes, it does not come down to earth in rain or snow. Most of it reaches the earth's surface as a result of molecular exchange reactions with the carbonates in seawater. A significant fraction of the C^{14} also is taken up by plants in the form of CO_2. Thereby it enters into the composition of all living substances. As soon as such a substance has become isolated from the biosphere the ratio of C^{14}/C^{12} begins to decrease. This ratio may therefore be used to establish the time of death. The isotope has found many applications; it has been employed for dating organic relics from historic and prehistoric periods, for establishing the date of the last ice age, and for measuring the age of the more recent deposits on the ocean floor. The usefulness of the C^{14} dating method is at present limited to events which occurred within the last 35,000 years.

Another useful isotope for obtaining information on geological events, particularly those which lie farther back in time, is Be^{10}, which has a half-life of 2.7 million years. It has been detected in the sediments of the deep Pacific where it slowly accumulates through millions of years, and it is now being used to establish a chronology of the ocean floor. At the same time, it may prove useful for establishing possible major changes in cosmic ray intensity during the last few million years.

The 12.5 year half-life of tritium makes this isotope well suited for a number of geophysical investigations. The vertical circulation of ocean water and the rate of exchange between ground water, inland lakes, the oceans, and the atmosphere have been investigated.

The remaining cosmic ray produced isotopes listed in Table 3 have been discovered comparatively recently. They are short-lived and their lifetime should make them valuable for studying the motion of air masses between various latitude belts, the exchange of air between the troposphere and different layers of the stratosphere, and for other meteorological problems which have a characteristic time scale of the order of weeks or months.

VII. Cosmic ray intensity in remote geological periods

It is difficult to find on the earth a permanent record of cosmic ray intensity in remote geological periods. No cosmic ray produced isotopes are known whose lifetime exceeds that of Be^{10} (2.7 million years), and while there are many types of stable nuclei produced in cosmic ray induced spallations, they cannot be distinguished from the very much larger number of similar isotopes already abundant on the earth. However, it is possible to learn something about the past cosmic ray activity in the planetary system by the study of meteorites. Meteorites show radiation damage in their crystal structure which is due to cosmic ray bombardment. This gives rise to the phenomenon of thermoluminescence, the emission of visible light upon heating. One also finds in meteorites not only long-lived cosmic ray produced isotopes such as tritium, Al^{26}, and Be^{10}, but also a number of stable elements, in particular, isotopes of the noble gases, helium, neon, and argon. These have an isotopic composition which differs markedly from that normally encountered and show clearly that the gases do not form an original constituent of meteoric matter but are cosmic ray produced. While the absolute amount of radioactive and stable cosmic ray products in a meteorite depends on the location of the sample with respect to the surface of the meteorite body, their ratio does not, and is therefore a measure of the amount of radiation which the whole (preatmospheric) piece has received since it was knocked off from a larger body in an interplanetary collision. If all meteorites had suffered fragmentation at the same time, the isotope ratios should be the same for all meteorites whose chemical composition is identical. Assuming an unchanged intensity of cosmic ray bombardment with time, one can define a so-called radiation age, which is proportional to the time which has elapsed since the meteorite was separated from its parent body and became subject to cosmic ray bombardment. Such ages have been determined and range from 10^7 to a few times 10^8 years; although experimental errors may still be large, it seems that different meteorites fragmentized at different periods. If the break-up times could be obtained by other means, the measurements would determine cosmic ray intensity in remote periods. At present this does not yet seem feasible.

It may be possible to make similar studies on micro-meteorites, which are presumably interstellar dust settling on the surface of the earth. This dust probably never formed part of a larger body, and therefore has been irradiated by cosmic radiation since the creation of our galaxy. The concentration of cosmic ray produced isotopes in micro-meteorites could, therefore, lead to information on the integrated intensity of cosmic radiation since the beginning of our galaxy.

While the past history of cosmic radiation is still largely unknown, it is nevertheless possible to state that some of the meteorite pieces which have been examined have received radiation doses equivalent to a present day cosmic ray intensity continued for a time of the order of 10^9 years. It seems possible that a detailed history of past cosmic ray intensity in the solar system will eventually be reconstructed; this will be of great interest because of the possible connection between past cosmic ray intensity and evolutionary changes on the earth, to which we have referred already.

VIII. New research problems

Although it is always hazardous to forecast future research developments, one can recognize three important technical advances made in 1957 which will greatly benefit cosmic ray research and will make it possible to attack some of the outstanding as yet unsolved problems. The three new technical features are: (a) the perfection of earth's satellites, (b) the organization for international cooperative cosmic ray observations through the International Geophysical Year (I.G.Y.), and (c) the establishment of permanent, well equipped, scientific stations in the Arctic and Antarctic.

At least two general classes of research problems can be specified which will be greatly assisted by these developments:

1. The radiation incident from outer space can be observed by means of satellites for long periods without interference from a layer of air which absorbs or converts much of the radiation. In particular, it will be possible to study the intensity and energy distribution of cosmic ray electrons; the charge and energy distribution of incident nuclei in the energy range between $\sim 10^3$ and 10^8 ev, which can reach the top of the atmosphere by spiralling along magnetic field lines entering at the poles; the γ-rays from the

decay of neutral pions produced in nuclear collisions of cosmic ray particles in interstellar space; and the possible emission of nuclear γ-rays from super novae and other active astrophysical objects.

2. The second class of problems deals with the observation of intensity variations in the cosmic radiation. Until now, cosmic ray work has been confined almost entirely to primary particles of energy $> 5 \times 10^8$ ev or their secondaries. In this energy region, time fluctuations are usually small, and major intensity variations extremely rare. Nevertheless, as discussed in Section IV, such fluctuations are very sensitive to changes in the distribution of fields and matter in the interplanetary space, and these in turn are largely governed by the emission of particles and of plasma clouds from the sun. Continuous observations of cosmic radiation, particularly in the very low energy band and the coordination of observations on a global scale, will no doubt prove to be an extremely effective method for studying conditions within the planetary system. The use of polar stations and of satellites, and the organization provided by IGY, will play a decisive role in these studies.

Also, in other fields, the importance of cosmic ray research is likely to increase, rather than diminish, in the next decade. The application of cosmic ray research to geophysics, and to the study of the geological past, is certain to gain in importance. In addition, cosmic ray physics will retain the monopoly which it enjoys at present for the study of ultrahigh energy particles (above a few hundred Bev).

Acknowledgment—A part of this article was written while the author was a guest at the University Institute for Theoretical Physics in Copenhagen. He wants to express his gratitude to the members of the Institute and in particular to thank Niels Bohr, Aage Bohr, and J. K. Bøggild for their hospitality.

Bibliography (principal references)

Much of the material covered in Sections II, III, IV, and V can be found in the series *Progress in Cosmic Ray Physics* (recently renamed *Progress in Elementary Particle and Cosmic Ray Physics*) published by the North Holland Publishing Co., Amsterdam. In particular, the following review articles in the series are relevant:

Section II: "The Nature of Primary Cosmic Radiation" by B. Peters, Vol. I, 1952.

Section III: "The Origin of Cosmic Radiation" by V. L. Ginzburg, Vol. IV, 1958.

Section IV: "Variations of Cosmic Ray Intensity" by H. Elliott, Vol. I, 1952.

Section V: "The Equilibrium of the Cosmic Ray Beam" by G. Puppi and N. Dallaporta, Vol. I, 1952.

and also "The Energy Balance of Cosmic Radiation" by G. Puppi, Vol. III, 1956.

Recent work on the composition of the primary radiation has been discussed by:

Appa Rao, M. V. K., S. Biswas, R. R. Daniel, K. A. Neelakantan and B. Peters, Abundance of light nuclei in the primary cosmic radiation, *Phys. Rev.*, 110, 751–765, 1958.

Important new results on the primary energy spectrum have been obtained by:

Clark, G., J. Earl, W. Kraushaar, J. Linsey, B. Rossi, and F. Scherb, An experiment on air showers produced by high-energy cosmic rays, *Nature* (London), 180, 406–409, 1957,

and by:

McDonald, F. B., Primary cosmic-ray proton and alpha flux near the geomagnetic equator, *Phys. Rev.*, 109, 1367–1375, 1958.

An up-to-date treatment of what is known of the relation between cosmic ray intensity and conditions in interplanetary space has been given in:

"Cosmic Ray Variations" by L. Dorman (1957), State Publishing House for Technical and Theoretical Literature, Moscow. (English Translation by Tech. Doc. Liaison Officer MCLTD, Wright Patterson Air Force Base, Ohio.)

No review articles have as yet appeared which cover the entire material of Sections VI or VII. Articles which deal with fairly broad aspects of the problems and can be consulted for additional references are:

Radio Carbon Dating, by W. F. Libby (1952), Univ. of Chicago Press.

"Bibliography of Radiocarbon Dating," *Quaternaria* II (1955) and IV (1957)

"On the Use of Tritium as a Tracer for Water in Nature" by B. Bolin (1958) in press.

"Cosmic Ray Produced Shortlived Radio Isotopes for Studying Large Scale Circulation of Air Masses" by B. Peters, *Jour. Terr. Atm. Phys.*, 13, 1958, in press.

"The Be^{10} Method for Studying Long Term Changes in Cosmic Radiation and the Chronology of the Ocean Floor". By B. Peters, *Z. Physik*, 148, 93, 1957.

"On the History of Meteorites as deduced from Isotope Analyses" by J. Geiss, *Chimia*, 11, 349, 1957.

(Manuscript received July 21, 1958.)

On the Response of Western Boundary Currents to Variable Wind Stresses

Takashi Ichiye

Oceanographic Institute
Florida State University
Tallahassee, Florida

Abstract—The forced response of ocean currents, particularly western boundary currents, to a variable system is analyzed. By dividing an ocean into a western zone and an interior region, a simple relation between wind stress torque over the whole ocean and mass transport of the western current can be obtained. Mass transport for the stationary case is compared with that derived from Stommel's frictional boundary theory. Then mass transport for wind stresses changing periodically is computed for two kinds of boundary conditions which lead to almost the same result. In a barotropic ocean or in a barotropic mode of a two-layered ocean, the amplitude of fluctuation of western mass transport approaches an equilibrium state with a phase lag of two to four days behind the wind when the period exceeds one month. On the other hand, amplitude and phase lag of fluctuation in the baroclinic mode are changing wildly, and approach the equilibrium state for a period longer than 10^3 days. The effect of non-linear inertia terms in the western current zone is negligible to the mass transport there. The divergence in the western zone causes waves moving southward and particularly the waves in the baroclinic mode to have a resonance period of about five months. Such a period seems to be associated with the observed predominance of half-year period fluctuation in the Gulf Stream mass transport.

1. *Introduction*—There are at present two principal theories concerning the nature of wind-driven ocean circulation: that of *Stommel* [1948] and *Munk* [1950]; and that of *Charney* [1955a] and *Morgan* [1956].

The first of these is based on the assumption that circulation is maintained by the balance between wind stresses, Coriolis' force, lateral or inner friction, and pressure force. In this theory, the western current zone is expressed as the frictional boundary layer, with the result that a larger eddy viscosity must be assumed to obtain a reasonable width for the western current than is required by the theory of Charney and Morgan. The existence and magnitude of such eddy viscosity require empirical demonstration, but no such evidence is yet available. Stommel and Munk's theory is expressed by convenient linear equations.

The second theory, proposed independently by Charney and Morgan, considers that the effect of lateral mixing is not necessary for balancing the wind torque. Instead, it is assumed that the non-linear inertia terms and β-term are predominant in the vorticity equation for balancing the wind stress curl in the western zone, while for the rest of the ocean a geostrophic relation holds. This might be termed the *inertia boundary theory*, since the western current is represented by a flow in the inertia boundary layer. This approach seems to be more realistic, since it is not dependent upon a hypothetical eddy viscosity. However, the equations introduced are non-linear and generally difficult to solve.

Of these two theories of stationary circulation, that of Stommel and Munk is the most convenient for application to a non-stationary case because of the linear equations involved. Such an application has already been attempted [*Ichiye*, 1951; *Veronis* and *Morgan*, 1955], although ambiguity about the eddy viscosity still remains.

Meanwhile, observations concerning the response of ocean circulation to a variable wind are meager, with data available only for a limited area and period. Such data may be derived from determination of mass transport by measuring the potential difference on both sides of a strong current by use of a submarine cable as across the Florida Strait [*Wertheim*, 1954; *Stommel*, 1957]. Frequent serial hydrographic surveys across a current such as the Gulf Stream [*Iselin*, 1940; *Von Arx*, 1955] constitute another means of deriving useful information on this problem. This latter method

might yield detailed data on the change in the western current, but the time lag in carrying out the surveys with existing instruments presents a serious handicap and the expense of operating a research vessel might be prohibitive.

On the other hand, the former method furnishes data on mass transport only, though it is possible to obtain continuous records for long periods of time comparatively inexpensively. It might be expected, therefore, that considerable data on mass transport may be accumulated in the future; and as a result, in the comparison of theory and observed data, an understanding of the relationships between a variable wind and mass transport of a western current should prove more important than knowledge of change in the minute structure of the current itself.

Thus, this paper attempts to derive the change of western mass transport caused by varying wind stresses. This can be done by dividing an ocean into a western boundary zone and an interior region as in the inertia boundary theory of Charney and Morgan; linear equations will thus be yielded, since non-linear terms then become negligibly small in the equation of mass transport in the western zone.

The response theory based on the viscous boundary theory [*Ichiye*, 1951; *Veronis* and *Morgan*, 1955] treated a homogeneous ocean, while the density stratification has been shown to play an important role in the response problem by *Charney* [1955b] and *Veronis* and *Stommel* [1956] who used a two-layered ocean model. However, their model has no boundary, so that the result is applied only to the response of a relatively short period such as is caused by a moving storm system. Motion with a longer period may be strongly influenced by coasts, and the proper boundary condition at the western coast should be non-linear.

The naive extension of this two-layer model to a bounded ocean by taking a linear-type boundary condition on both coasts encounters a serious paradox in that the amplitude of induced Rossby waves fluctuates wildly for a change of the period of wind, as pointed out by *Munk* [1957, unpublished manuscript]. This situation was improved by introducing a friction force with a reasonable order of friction coefficient which damps out the fluctuating amplitudes of the Rossby waves [*Ichiye*, 1958]. However, the solution thus obtained is also much influenced by choice of the friction coefficient, the reasonable estimation of which is possible only by comparison with observed data, which are not easily obtained at present as stated above.

It is possible to obtain a very simple relation between the change in mass transport of the western current and that of wind stresses even in a two-layered ocean, and a remarkable difference between barotropic and baroclinic modes is shown in such a model.

2. *Basic equations of the wind-driven current in a two-layer ocean*—First of all, we deduce equations of motion and continuity in a two-layered ocean, and afterwards we may simplify them in the case of a homogeneous ocean. In such an ocean, we can consider that the wind stress acts as a body force in the upper layer only. Further, we divide the ocean into an interior region and a western current zone, where the former is represented by a rectangle from $x = 0$ to $x = a$ and the latter by a narrow strip from $x = -d$ to $x = 0$ in a cartesian coordinate with x and y directed eastwards and northwards, respectively, as shown in Figure 1.

In the interior region the equations of motion and continuity are given by *Charney* [1955b], and *Ichiye* [1958]

$$\frac{\partial u}{\partial t} - fv = -g\frac{\partial(\eta + \eta')}{\partial x} + \frac{\tau_x}{H} \tag{2.1}$$

$$\frac{\partial v}{\partial t} + fu = -g\frac{\partial(\eta + \eta')}{\partial y} + \frac{\tau_y}{H} \tag{2.2}$$

$$\frac{\partial \eta}{\partial t} + H\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right) = 0 \tag{2.3}$$

$$\frac{\partial u'}{\partial t} - fv' = -g\frac{\partial(\epsilon\eta + \eta')}{\partial x} \tag{2.4}$$

$$\frac{\partial v'}{\partial t} + fu' = -g\frac{\partial(\epsilon\eta + \eta')}{\partial y} \tag{2.5}$$

$$\frac{\partial \eta'}{\partial t} + H'\left(\frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y}\right) = 0 \tag{2.6}$$

where the symbols with prime denote corresponding quantities in the lower layer and H is the equilibrium thickness of the upper layer, η is the deviation of H, τ_x and τ_y are wind stress components ϵ is the density ratio ρ/ρ'.

On the other hand, in the western current zone we should retain the inertia terms caused

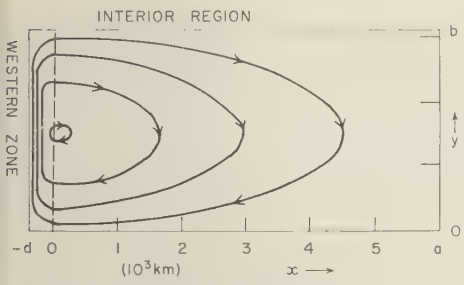


FIG. 1—Geometrical arrangement of a model ocean and schematic stream lines ($a = 6 \times 10^3$, $b = \pi \times 10^3$, $d = 4 \times 10^2$ *km*)

by a rapid variation in x direction from the western coast to the boundary between interior and western zone $x = 0$. Then the equation of motion in x direction in the upper layer becomes

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial u^2}{\partial x} - fv = -g\frac{\partial(\eta + \eta')}{\partial x} \qquad (2.1')$$

instead of (2.1). This type of equation will be discussed in section 7 and in the following three sections we use the equation of continuity only in order to deduce the mass transport of the western current by neglecting the divergence term.

When we take an average of continuity equations (2.3) and (2.6) over the width of the western zone, we obtain easily

$$u_0/d + \partial\bar{v}/\partial y = 0 \qquad (2.7)$$

$$u_0'/d + \partial\bar{v}'/\partial y = 0 \qquad (2.8)$$

in which u_0 and u_0' are values of u and u' at $x = 0$ and

$$\bar{v} = \frac{1}{d}\int_{-d}^{0} v\,dx, \quad \bar{v}' = \frac{1}{d}\int_{-d}^{0} v'\,dx$$

since u and u' should vanish at the coast $x = -d$.

3. *Stationary circulation in a homogeneous ocean*—First, it is reasonable to compare the present model with the solution of *Stommel* [1948] in a rectangular homogeneous ocean with a frictional force in a steady state.

In this case we have only to use equations (2.1), (2.2), and (2.3) with $\eta' = 0$, $\partial\eta/\partial t = 0$. Then the vorticity equation in the interior region is given by

$$\beta v = \text{curl}\ \tau/H \qquad (3.1)$$

Further, there is no divergence; so that u and v have a velocity potential

$$v = \partial\psi/\partial x, \quad u = -\partial\psi/\partial y$$

Therefore we have at once

$$\psi = \frac{\beta}{H}\int_a^x \text{curl}\ \tau\,dx \qquad (3.2)$$

under the boundary condition at the eastern coast $x = a$ such as $u = 0$.

Then the mass transport equation in the western zone (2.7) becomes simply

$$\partial\bar{v}/\partial y = -(\partial\psi/\partial y)_{x=0}/d$$

or when we assume that $\bar{v} = 0$ at $y = 0$, we have

$$\bar{v} = \{\psi(0, 0) - \psi(0, y)\}/d \qquad (3.3)$$

Following *Stommel* [1948], we take the stress distribution

$$\tau_x = -\tau_0 \cos ly, \quad \tau_y = 0, \quad l = \pi/b \qquad (3.4)$$

with the value of y from 0 to b, representing a westerly wind in the northern latitude and a trade wind in the low latitude. Then, we can obtain

$$dH\bar{v} = \frac{al}{\beta}\tau_0 \sin ly \qquad (3.5)$$

as a total mass transport of the western current.

On the other hand, the original theory of *Stommel* [1948] gives a vorticity equation

$$k\Delta\psi + \beta\psi_x = \text{curl}\ \tau/H \qquad (3.6)$$

in which k is a friction coefficient. The solution under the boundary conditions that $u = 0$ at both coasts $x = a$ and $x = -d$ is given by

$$\psi = \frac{\tau_0}{kl}(1 - Ae^{\alpha_1\overline{x+d}} - Be^{\alpha_2\overline{x+d}})\sin ly \qquad (3.7)$$

in which

$$\alpha_{1,2} = -p \pm q, \quad p = \beta/2k, \quad q^2 = p^2 + l^2 \qquad (3.8)$$

$$A = (1 - e^{\alpha_2 a'})/(e^{\alpha_1 a'} - e^{\alpha_2 a'}), \quad B = 1 - A, \quad a' = a + d \qquad (3.9)$$

For a reasonable range of k less than 2×10^{-5} (sec^{-1}) and values of constants such as

$$a = 6 \times 10^8 \text{ cm}, \quad d = 4 \times 10^7 \text{ cm}, \quad b = \pi \times 10^8 \text{ cm} \tag{3.10}$$

we can put $|e^{\alpha_2 a'}| \ll 1$. (The range of k is estimated from the relation given by *Stommel* [1948], $k = 10^{-2}/H$.) Then the mean value of v from $x = -d$ to $x = 0$ is given by

$$\bar{v} = \frac{1}{d}\int_{-d}^{0} \frac{\partial \psi}{\partial x}\, dx = \frac{1}{d}\,\psi(0, y) \tag{3.11}$$

since $\psi(-d, y) = 0$ from the boundary condition.

Therefore a total mass transport in the western zone corresponding to (3.5) is given by

$$dH\bar{v} = R_m \frac{al}{\beta}\,\tau_0 \sin ly, \tag{3.12}$$

$$R_m = \frac{\beta}{akl^2}\{1 - e^{\alpha_1 a} - (1 - e^{\alpha_1 a'})e^{\alpha_2 d}\}$$

under the condition $|e^{\alpha_2 a'}| \ll 1$. The ratio R_m indicates how the present simple model approximates Stommel's exact solution of the frictional boundary layer theory.

In Figure 2 we indicate this ratio as a function

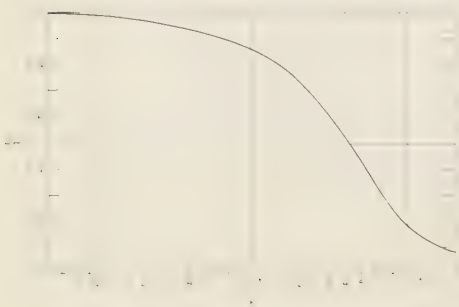


FIG. 2—The ratio R_m as a function of k

of k for the range of $2 \times 10^{-5} \geq k \geq 10^{-7}$. This curve shows that the ratio is approaching unity as k decreases and that the present model approximates Stommel's solution within an error of 10 pct for k less than 7×10^{-7}, which corresponds to the effective depth of circulation of 280 m or more. When friction increases, the width of the western current increases and thus the mass transport within the western zone of finite width b decreases. This is the reason why the ratio decreases with increasing k. However, there is a good reason why k is less than 10^{-7} in the interior region from the estimation of *Stommel* [1948] and also from *Morgan*'s discussion [1956]; therefore the simplified treatment here can be used with reasonable accuracy for computation of the western mass transport.

In the above discussion we can use geostrophic relationships in the vorticity equation (3.1) such as

$$u = -\frac{g}{f}\frac{\partial \eta}{\partial y}, \quad v = \frac{g}{f}\frac{\partial \eta}{\partial x} \tag{3.13}$$

instead of the velocity potential. We can integrate the vorticity equation for η and have

$$\eta = \frac{f}{\beta g H}\int_a^x \text{curl } \tau\, dx \tag{3.14}$$

We obtain the value of u in the interior region by putting it into (3.14) which leads to the same result as before if we treat f as a constant in the first equation of (3.13). The validity of the geostrophic approximation in the interior of the ocean has been proved by *Charney* [1955ab] and *Ichiye* [1958] for a variable wind stress of a relatively long period in which the gravity and inertia waves are negligible. Such a situation makes the subsequent analysis simple and clear.

4. *The response of a homogeneous ocean (with one Rossby wave)*—Now we consider the response of a homogeneous ocean to wind stresses changing periodically with time. By cross-differentiation of (2.1) and (2.2) (with $\eta' = 0$) we have the vorticity equation such as

$$\frac{\partial \zeta}{\partial t} - \frac{f}{H}\frac{\partial \eta}{\partial t} + \beta v = \frac{1}{H}\,\text{curl } \tau \tag{4.1}$$

Substituting the geostrophic relation (3.13), we have

$$\frac{\partial}{\partial t}(\Delta\eta) - \frac{1}{\lambda^2}\frac{\partial \eta}{\partial t} + \beta\frac{\partial \eta}{\partial x} = \frac{f}{gH}\,\text{curl } \tau \tag{4.2}$$

$$\lambda^2 = gH/f^2$$

where Δ is the Laplacian. When we assume the same spatial distribution of wind stresses as (3.4) with a factor changing periodically with time such as

$$\tau_x = -\tau_0 \cos ly \cos \omega t, \quad \tau_y = 0 \tag{4.3}$$

then the solution of (4.2) is expressed by

$$\eta = \mathbf{R}[p(x)e^{i\omega t}] \sin ly \tag{4.4}$$

where

$$P(x) = -iP_0(1 + Ae^{\alpha_1 x i} + Be^{\alpha_2 x i}) \qquad (4.5)$$

with

$$P_0 = F/m^2, \quad F = fl\tau_0/gH, \quad m^2 = l^2 + \lambda^{-2}$$

and two parameters α_1 and α_2 are given by

$$\alpha_1 = p + q, \quad \alpha_2 = p - q \qquad (4.6)$$

with

$$p = \beta/2\omega, \quad q = \{(\beta/2\omega)^2 - m^2\}^{-\frac{1}{2}}$$

These two terms with $e^{\alpha_j x i}$ correspond to two Rossby waves, both of which have a westerly phase velocity, but in which the shorter wave (α_1-wave) has a group velocity eastward, while the longer wave (α_2-wave) has a group velocity westward [*Arons* and *Stommel*, 1956].

The amplitudes A and B of these waves are determined by the boundary conditions [*Ichiye*, 1958]. It is obvious that the condition at the eastern coast is that the velocity vanishes; this leads to a linear relation $\partial\eta/\partial y = 0$ or $P(a) = 0$. In order to specify two unknown constants A and B, we need another condition, and we have two alternatives for such a condition.

One of these is the elimination of the α_1 wave in the interior region *a priori*. This is reasonable because the wave of a shorter wave length becomes almost negligible for a period longer than a few days in the interior region, when we take into account a small friction force, as shown by *Ichiye* [1958]. If we neglect friction, a naive application of the boundary condition such as vanishment of u at the western coast leads to a paradoxical result of wildly oscillating amplitudes in the limiting case of $\omega \to 0$, as discussed by *Munk* [1957, unpublished report]. On the other hand, when we put $A = 0$ and determine B by the condition that

$$P(a) = 0 \qquad (4.7)$$

we have

$$P(x) = P_0[1 - e^{\alpha_2(x-a)i}] \qquad (4.8)$$

or substituting this into (4.4) we have an equation for a western mass transport similar to (3.5)

$$dH\bar{v} = \frac{l\tau_0}{m^2\omega}\{\sin\omega t - \sin(\omega t - \alpha_2 a)\}\sin ly \qquad (4.9)$$

For a small value of ω we can put

$$q \doteq \frac{\beta}{2\omega} - \frac{\omega}{\beta}m^2 \quad \text{or} \quad \alpha_2 \doteq \frac{\omega}{\beta}m^2 \qquad (4.10)$$

and thus the equation (4.9) approaches

$$dH\bar{v} = \frac{l\tau_0}{\beta}a\sin ly\cos\omega t \qquad (4.11)$$

since

$$\sin(\omega t - \alpha_2 a) \to \sin\omega t - (a\omega m^2/\beta)\cos\omega t \qquad (4.12)$$

under the condition that

$$a\omega m^2\beta \ll 1$$

The equation (4.11) corresponds to the solution (3.5) in the stationary case. Thus we can prove that the solution (4.9) is asymptotic to the equation in the stationary case as $\omega \to 0$. When we give the following numerical values for constants

$$\beta = 2 \times 10^{-13}, \quad H = 10^5,$$
$$f = 10^{-4}, \quad a = 6 \times 10^8, \quad l = 10^{-8} \qquad (4.13)$$

then we have

$$\lambda^{-2} = 10^{-8}, \quad m^2 = 2 \times 10^{-8} \qquad (4.14)$$

so that, for example, for $\omega \leq 10^{-6}$ or for a period longer than 72.8 days, we have $(\beta/2\omega) \geq 10^{-7}$, and thus the approximation (4.10) is valid, with a value of $a\omega m^2/\beta = 0.6$; then we may expect that the approximation (4.11) could be attained for a period of a few years. In order to see for what period T the equation (4.9) approaches the stationary solution (4.11), we compute the amplitude ratio of (4.9) to (4.11) r and the phase lag of the former equation behind the latter θ as a function of T from the equations such as

$$r = \frac{\beta}{a\omega m^2}\{(1 - \cos\alpha_2 a)^2 + \sin^2\alpha_2 a\}^{\frac{1}{2}} \qquad (4.15)$$

$$\tan\theta = (1 - \cos\alpha_2 a)/\sin\alpha_2 a \qquad (4.16)$$

and the result is shown in Figure 3.

Particularly when $\omega \geq 1/\sqrt{2} \times 10^{-5}$ or a period shorter than 10.3 days, the value of q becomes imaginary, so that the solution corresponding to (4.9) becomes

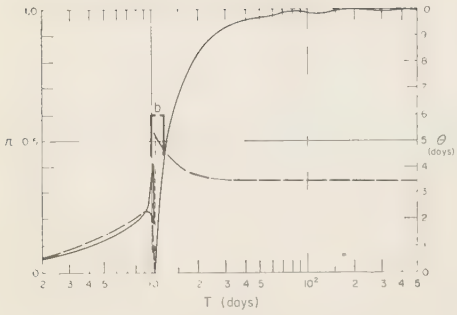


FIG. 3a—r (full line) and θ (in days, broken line) as a function of period T (in days) for one type of Rossby waves. (The parts of curves indicated by "b" are also shown in Fig. 3b in details. The dotted line indicates the gap of θ by π)

$$dH\bar{v} = \frac{l\tau_0}{m^2\omega}\{\sin\omega t - e^{-q'a}\sin(\omega t - pa)\}, \qquad (q')^2 = m^2 - p^2 \qquad (4.17)$$

and the ratio r and phase lag θ should be computed by use of (4.17) instead of (4.9). Near this period the ratio and phase lag change quite conspicuously as shown in Figure 3b.

We can see immediately that the amplitude of transient model (4.9) is asymptotic to that of a stationary model of a period of about 50 days or longer, for example, and also that phase lag is an almost constant value of about 3.5 days for a period longer than approximately 20 days. The rapid change near the critical value of ω is seen both in the amplitude ratio and phase lag. The maximum ratio is attained at the critical frequency $\omega_c = 1/\sqrt{2} \times 10^{-5}$ or T_c (critical period). Also the phase lag θ has a gap of π at about $T = 10.7$ days during which the amplitude of the response vanishes.

When we call the range of period less than T_c *SC* (Sub-critical), after *Ichiye* [1958], we can see that in the *SC* range the α_2-waves (in *Ichiye* [1958] indicated as δ_2-wave) decrease their amplitude very rapidly as they depart from the east coast, since they have a factor such as $e^{-q'(a-x)}$, and the second term of (4.17) becomes negligible except very near T_c. Thus we have $\theta = \pi/2$ and r decreases proportional to $1/T$. However, in such *SC* range, the effect of inertia and gravity waves becomes predominant for a period shorter than one to two days, and we cannot apply the conclusion here to phenomena with such a short period.

5. *The response of a homogeneous ocean (with two Rossby waves)*—As an alternative condition for determining the amplitude of two Rossby waves, we may take the boundary condition at $x = 0$. In the actual situation, this boundary is not a fixed one, and the condition there should be non-linear if we wish to know the change of detailed structure of the western current. However, we are only concerned with the total mass transport here and we can take a very simple condition without damaging the whole system of the solution.

When we look at the stream lines of the stationary solution given by *Stommel* [1948], we see that they are almost parallel to the x axis as shown in Figure 1. Therefore, it is quite reasonable to assume the boundary condition

$$x = 0, \quad v = 0, \quad \text{or} \quad \partial\eta/\partial x = 0 \qquad (5.1)$$

In other words there is no parallel component of velocity to the boundary. Therefore we exclude the western boundary eddy, such as the Sargasso Sea which appeared in the solution of the frictional boundary layer of *Munk* [1950], because in this model we neglect the lateral eddy viscosity in the interior region. However, this is not so serious as might be supposed, since we can consider that such a boundary eddy is included in the boundary region between $x = -d$ to

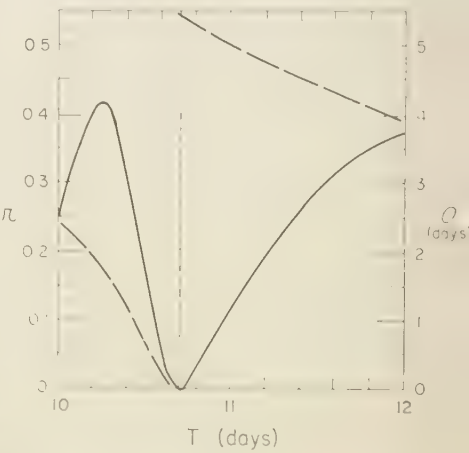


FIG. 3b—Detailed curves of r (full line) and θ (broken line)

$x = 0$, owing to the condition that the contribution from such eddy is only a small fraction of the mass transport of the western current.

Then from the condition (4.7) with (5.1) we have

$$A = \frac{p - q}{2} \cdot \frac{e^{-apl}}{q \cos aq - ip \sin aq}$$

$$B = -\frac{p + q}{2} \cdot \frac{e^{-apl}}{q \cos aq - ip \sin aq} \qquad (5.2)$$

Putting

$$s^2 = (q \cos aq)^2 + (p \sin aq)^2$$

$$\tan \phi = (p \tan aq)/q \qquad (5.3)$$

we can get the equation for $P(x)$ as

$$P(x) = -P_0 i \left\{ 1 + \frac{p-q}{2s} \exp i(\overline{p+q}x - ap + \phi) - \frac{p+q}{2s} \exp i(\overline{p-q}x - ap + \phi) \right\} \qquad (5.4)$$

in which

$$P_0 = (fl\tau_0)/(m^2 \omega g H) \qquad (5.5)$$

as in (4.5).

Thus, similar to (4.9), we have

$$dH\bar{v} = \frac{l\tau_0}{m^2\omega} \{\sin \omega t - \frac{q}{s} \sin (\omega t - ap + \phi)\} \qquad (5.6)$$

When $T < T_c$, the value of q becomes imaginary and (5.4) and (5.6) are expressed by substituting q with $q' = (m^2 - p^2)^{1/2}$ and s, ϕ of (5.3) with s', ϕ'

$$(s')^2 = (q')^2 \cosh^2 aq' + p^2 \sinh^2 aq'$$

$$\tan \phi' = (p \tanh aq')/q' \qquad (5.7)$$

This solution also leads to the stationary solution of (4.11) when $\omega \to 0$ because from (5.3) we have

$$p/q \to 1, \quad q/s \to 1,$$

or

$$\tan \phi \to \tan aq \quad \text{as} \quad \omega \to 0$$

and thus

$$\phi \to aq$$

which gives

$$ap - \phi \approx m^2\omega/\beta$$

and therefore the same result as in the previous section.

Corresponding to equations (4.15) and (4.16) we have

$$r^2 = \left(1 - \frac{q}{s} \cos \overline{ap - \phi}\right)^2 + \frac{q^2}{s^2} \sin^2 (ap - \phi) \qquad (5.8)$$

$$\tan \theta = \left(1 - \frac{q}{s} \cos \overline{ap - \phi}\right) \Big/ \sin (ap - \phi) \qquad (5.9)$$

where in the *SC* range we should replace q, s and ϕ with q', s' and ϕ' as given by (5.7).

We compute the change of r and θ with a period T longer than two days, and the result is shown in Figure 4 with a similar representation to Figure 3, having detailed curves near the critical period of $T_c = 10.3$ days. The response curves show almost the same tendency as is shown in Figure 3 in both the amplitude ratio and phase lag for ranges of period longer than 50 days and shorter than 10 days. In the intermediate range of period, both amplitude and phase curves indicate some unevenness which is caused by co-existence of α_1-wave of shorter wave length. Further, in the neighborhood of T_c, the response characteristics are slightly different from those in the previous case. However, such differences do not seem to be important, because in applying the theory to a real situation in the ocean it is difficult to

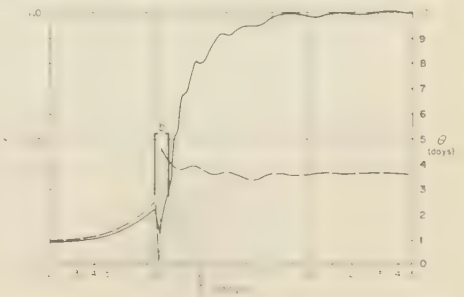


FIG. 4a—r (full line) and θ (broken line) for two types of Rossby waves (see the caption of Fig. 3a)

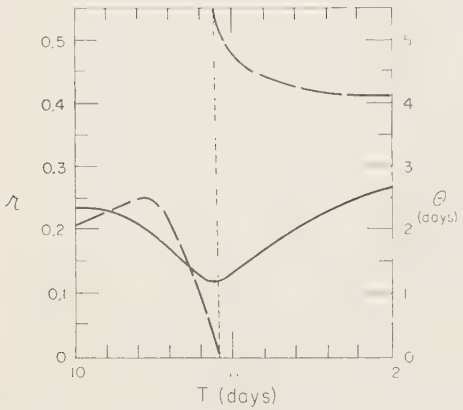


FIG. 4b—Detailed curves of r (full line) and θ (broken line)

detect these differences from the observation and thus we can hardly decide which condition in two alternatives is more proper in a real ocean.

On the other hand, from a theoretical point of view, it is interesting that the condition present at the western boundary (not the western coast) leads to a solution which approaches the stationary type as $\omega \to 0$ and also avoids a paradoxical result of wildly fluctuating amplitudes as in case of the condition that $u = 0$ at the coast.

6. *The response of a two-layer ocean*—As in a homogeneous ocean, we can obtain the vorticity equation for the solution in the model of a two-layered ocean by cross-differentiation of (2.1), (2.2), and (2.4), (2.5) for the upper and lower layer, respectively. Then we have

$$\nabla^2 \frac{\partial}{\partial t}(\eta + \eta') - \lambda^{-2} \frac{\partial \eta}{\partial t} + \beta \frac{\partial}{\partial x}(\eta + \eta') = \frac{f}{gH} \operatorname{curl} \tau \tag{6.1}$$

$$\nabla^2 \frac{\partial}{\partial t}(\epsilon\eta + \eta') - (\lambda')^{-2} \frac{\partial \eta}{\partial t} + \beta \frac{\partial}{\partial x}(\epsilon\eta + \eta') = 0 \tag{6.2}$$

Under a geostrophic assumption (3.13) [*Charney*, 1955b; *Ichiye*, 1958] where

$$\lambda^2 = gH/f^2, \quad (\lambda')^2 = gH'/f^2$$

In order to integrate these two equations we can apply the method of separation of motion into two normal modes, barotropic and baroclinic [*Charney*, 1955b; *Veronis* and *Stommel*, 1956; *Ichiye*, 1958], because in this model boundary conditions are linear. Then from (6.1) and (6.2) we have

$$\nabla^2 \frac{\partial R_j}{\partial t} - \lambda_j^{-2} \frac{\partial R_j}{\partial t} + \beta \frac{\partial R_j}{\partial x} = \sigma_j \frac{f}{gH} \operatorname{curl} \tau \tag{6.3}$$

in which

$$R_j = \sigma_j(\eta + \eta') + \epsilon\eta + \eta', \quad (j = r, l) \tag{6.4}$$

and suffixes r and l mean barotropic and baroclinic mode, respectively. The constants σ_r and σ_l are given by

$$\sigma_r = r - (1 - \epsilon)r/(1 + r) \approx r, \qquad \sigma_l = -1 + (1 - \epsilon)r/(1 + r) \approx -1 \tag{6.5}$$

with

$$r = H/H'$$

and also

$$\lambda_j^{-2} = (\kappa_j/\lambda)^2$$

with

$$\kappa_r^2 \approx r/(1 + r), \quad \kappa_l^2 \approx (1 + r)/(1 - \epsilon)$$

Thus

$$\lambda_r^2 \approx g(H + H')/f^2, \qquad \lambda_l^2 \approx g(1 - \epsilon)H/\epsilon f^2 \tag{6.6}$$

We can express pressure terms in (6.1) and (6.2) by next relations

$$\eta + \eta' \approx (R_r - R_l)/(1 + r), \qquad \epsilon\eta + \eta' \approx (R_r + rR_l)/(1 + r) \tag{6.7}$$

The solution of (6.3) corresponding to the stress distribution (4.3) is given by

$$R_j = -\frac{\sigma_j l f \tau_0 i}{\omega m_j^2 gH}(1 + A_j e^{\alpha_{1j} z i} + B_j e^{\alpha_{2j} z i}) \tag{6.8}$$

where

$$m_j^2 = l^2 + \lambda_j^{-2}, \qquad \alpha_{1j} = p + (p^2 - m_j^2)^{\frac{1}{2}}, \qquad \alpha_{2j} = p - (p^2 - m_j^2)^{\frac{1}{2}} \tag{6.9}$$

and A_j and B_j are determined from the boundary condition at $x = a$ and two alternate conditions as in the homogeneous ocean.

The mass transport of the western current can be obtained by the same process as before. First, in order to get the solution for an equilibrium case, we put the terms with $\partial/\partial t$ in (6.3) to zero and we have

$$R_j = \frac{\sigma_j f l(a - x)}{\beta g H} \cos \omega t \sin ly \qquad (6.10)$$

Then the mass transport of upper and lower layer is given by

$$dH\bar{v} = \frac{gH}{f} (\eta + \eta')_{x=0} \approx \frac{al\tau_0}{\beta} \sin ly \cos \omega t \qquad (6.11)$$

and

$$dH'\overline{v'} = \frac{gH'}{f} (\epsilon\eta + \eta')_{x=0} \approx 0 \qquad (6.12)$$

Therefore, there is no mass transport in the lower layer in a stationary or equilibrium state, as is easily seen from the original equation (6.2).

Generally in a transitional case we can express the mass transport of each layer by use of the third term or (6.11) or (6.12) with the value of R_j at $x = 0$. Thus, to show the response of each layer, we should express R_{j0} (value at $x = 0$) by parameters, similar to r and θ of (4.15), (4.16) or (5.8), (5.9) such as

$$(gHR_{j0})/f = r_j M \cos(\omega t - \theta_j) \sin ly \qquad (6.13)$$

where $M = al\tau_0/\beta$ is the amplitude of the total mass transport in the equilibrium state given by (6.11). Further, r_j is the amplitude ratio of each mode to that of the total mass transport of the equilibrium state, and θ_j is the phase lag behind it.

When we assume the condition (5.1) we have

$$(gHR_{j0})/(fM) = \frac{\beta\sigma_j}{a\omega m_j^2} \left\{ \sin \omega t - \frac{q_j}{s_j} \sin(\omega t - ap + \phi_j) \right\} \qquad (6.14)$$

where q_j, s_j, and ϕ_j are obtained by replacing m with m_j. When we take $A_j = 0$ as the second condition, as in section 4, then we have only to put $q_j/s_j = 1$ and $\phi_j = aq_j$ in this equation.

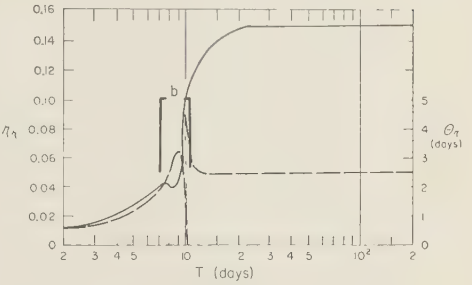


FIG. 5a—r (full line) and θ (broken line) for a barotropic mode of a two-layered ocean (See the caption of Fig. 3a)

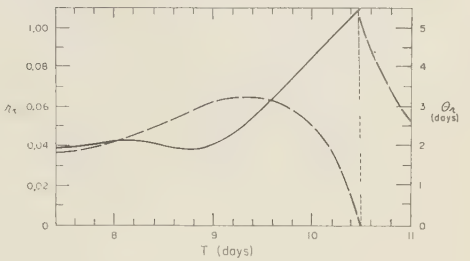


FIG. 5b—Detailed curves of r (full line) and θ (broken line)

In Figure 5 we show the curves of r_r and θ_r from (6.14) for the second condition (5.1) as functions of period T with detailed features near $T_e = 8.15$ days as in Figures 3 and 5, by assuming $H = 600m$, $H' = 3400m$. For the second condition of $A_r = 0$, the general curve of Figure 6 can be applied except in a range of T near $T = T_e$; therefore the curve is not shown here. The general tendency of these curves is quite similar to that shown in Figure 4, as expected, since they are the response curves of a barotropic mode, with the only difference of depth assumed here being $H + H' = 4000m$. Thus the critical period in this model is shorter than that of Figure 5. Further, the limiting value of r_r as $\omega \to 0$ is $\sigma_r \approx 0.15$ instead of one in the previous model.

On the other hand, the response curves of a baroclinic mode such as r_l and θ_l are quite different from those obtained before. In Figure 6 we show the curves of $-r_l$ and θ_l as functions of T in two ranges of period, that is, in the neighborhood of $T_e = 210$ days and for longer period,

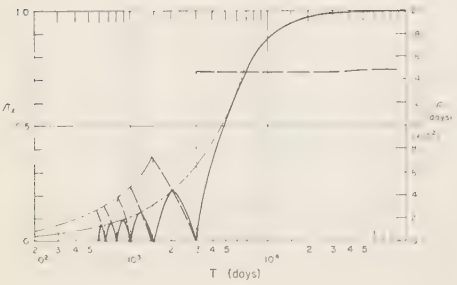


FIG. 6a—r (full line) and θ (broken line) for a baroclinic mode of a two-layered ocean. (The chain lines indicate the envelope curves equal to $\beta T/a\pi m_r^2$ for r and equal to $T/2$ for θ, respectively. See the caption of Fig. 3a)

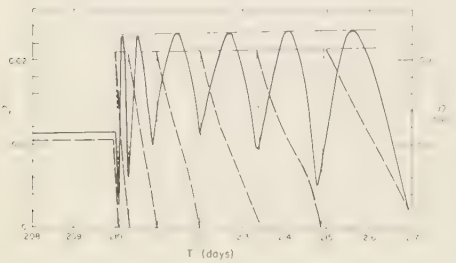


FIG. 6b—Detailed curves of r (full line) and θ (broken line) near T_c for one type of Rossby waves. (See the captions of Fig. 6a and Fig. 3a)

assuming the second condition of (5.1). In Figure 7, we show the curves of $-r_l$ and θ_l for the second condition that $A_l = 0$ for a range of period near T_c, since the response curves for longer periods are almost the same as those shown in Figure 8.

The conspicuous feature of the response of the baroclinic mode is that the amplitude and phase lag of the response become stationary for a period longer than ten days, in which case the ratio $-r_l$ approaches to 1 and the phase lag to 1.5×10^3 days. Between this range of period and T_c, the amplitude ratio indicates spectral distribution such that at some nodal periods the ratio r_l vanishes and the phase lag has a gap of π. The interval between successive nodal periods increases as the period increases from the critical period T_c, so that near T_c the interval is quite short, a fraction of a day, as shown in both figures of Figures 6 and 7. In the SC range or for $T < T_c$, the equation (6.14) approaches

$$-r_l = \frac{\beta}{a\omega m_l^2} \sin \omega t \qquad (6.15)$$

$$= 5.51 \times 10^{-5} T \sin \omega t, \quad (T \text{ in days})$$

quite rapidly as the period decreases from T_c, because of exponential factor in exp $(\alpha_l xi)$, as stated in the previous sections.

From (6.11) and (6.12) the transport in the upper and lower layer are expressed by

$$N = M_r \cos(\omega t - \theta_r) - M_l \cos(\omega t - \theta_l) \qquad (6.16)$$

$$N' = M_r \cos(\omega t - \theta_r)/r + M_l \cos(\omega t - \theta_l) \qquad (6.17)$$

in which

$$M_j = (r_j a l \tau_0)/\beta(1 + r), \quad (j = r, l)$$

with a minus sign in r_l. As shown in Figures 5 and 6, $\theta_r \approx 2.5$ days for a period longer than about 15 days; therefore we can consider that the barotropic mode has almost no phase lag for a period longer than 100 days.

When we consider a synthetic picture of both modes as a whole, we can conclude that the barotropic mode is predominant in the range of a period less than $T_c = 210$ days, since its amplitude is ten or more times that of the baroclinic mode. Both modes have the same order of amplitude in a range of a period of 10^3 to 2×10^3 days, although in this range the baroclinic mode shows the nodal period as seen in Figure 6. For a period longer than 10^4 days the baroclinic mode is predominant; thus the transport of the western current is contributed by the upper layer only.

Owing to the spectral distribution of the amplitude ratio between T_c and 10^4 days in a baroclinic mode, there is a very interesting state concerning the mode of a fluctuation of the western current. For instance, we can see from Figure 6 that r_l is almost of the same order as or larger than r_r for a range of period between 1.7×10^3 and 2.5×10^3 days; so that in such a range of period the velocity in the upper layer in the western zone is greater than in the lower layer with a phase lag behind the wind of 200 days or more as estimated from (6.16). The lower layer transport of the water is, instead, almost without a phase lag, since in this range

θ_r is only about 2.5 days. On the other hand, at a period of 1.4×10^3 or 3.0×10^3 days, $r_l = 0$ or $M_l = 0$ and the upper and lower layer have an equal velocity without a phase lag to the wind system.

7. *Effects of inertia terms and divergence in the western zone*—Heretofore we have neglected a divergence in the western zone and thus obtained the mean value $\bar{v}$ in this zone from the equation of continuity only. If we keep the divergence, we have to use the equation of motion in the western zone in order to obtain $\bar{v}$. Then this equation should be given by (2.1′), since in this zone the gradients of u and v in x-direction have large values. On the other hand, we can prove that the inertia terms become negligible when we consider the averaged current $\bar{u}$, $\bar{v}$ instead of the individual value.

In a two-layered ocean, the equations of motion and continuity (2.1′) and (2.3) in the western zone are integrated with respect to x from $x = -d$ to $x = 0$ and are divided by d. Then we have

$$\frac{\partial}{\partial t}\bar{u} + \frac{1}{2d}u_0^2 - f\bar{v} = -\frac{g}{d}(\xi_0 - \xi_b) \qquad (7.1)$$

$$\frac{u_0}{d} + \frac{\partial \bar{v}}{\partial y} = -\frac{1}{H}\frac{\partial \bar{\eta}}{\partial t} \qquad (7.2)$$

with corresponding equations in the lower layer, in which

$$\xi = \eta + \eta', \quad \xi' = \epsilon\eta + \eta'$$

The subscript 0 means the value at $x = 0$; subscript b means that of the coast $x = -d$; and the bar indicates the average between $x = -d$ and $x = 0$. In derivation of these equations we use the boundary condition that $u_b = 0$. In them $\bar{u}$, $\bar{u}'$, $\bar{\eta}$ and $\bar{\eta}'$ are unknown, but they are assumed to be given by the average of their values at $x = 0$ and $-d$. Then we can put

$$\bar{\eta} = \gamma_1(\eta_0 + \eta_b)/2, \quad \bar{u} = \gamma_2 u_0/2 \qquad (7.3)$$

with corresponding relations for $\bar{\eta}'$ and $\bar{u}'$, in which γ_i and γ_i' are constants having values close to 1.

Then η_0 and η_0' are given by the solutions of R_{j0} in the interior region and u_0 and u_0' are expressed by R_{j0} from the geostrophic approximation. Therefore $\bar{v}$, $\bar{v}'$, η_b, and η_b' are unknown quantities which can be obtained by solving

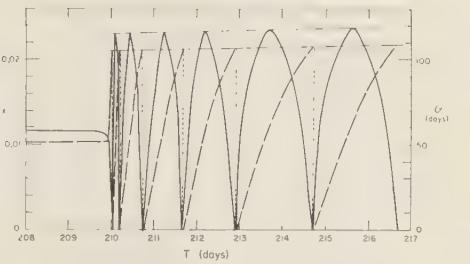


FIG. 7—Detailed curves of r (full line) and θ (broken line) near T_e for two types of Rossby waves. (See the captions of Fig. 6a and Fig. 3a)

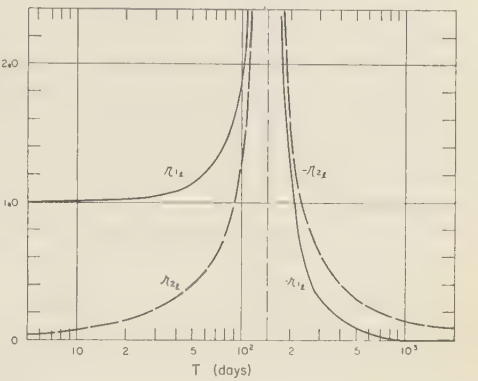


FIG. 8—The coefficients $|r_1|$ (full line) and $|r_2|$ (broken line) for a baroclinic mode

four equations, (7.1), (7.2) and corresponding equations in the lower layer.

First, it is more convenient to estimate the order of inertia terms in the equation (7.1). When we assume that

$$\frac{\partial}{\partial t} = \omega \leq 10^{-5}, \quad \frac{\partial}{\partial y} = l \approx 10^{-8}$$

the inertia terms are estimated as

$$\left|\frac{d}{2g}\frac{\partial \bar{u}}{\partial t}\right| = \left|\frac{d\gamma_2}{2f}\frac{\partial^2 \xi_0}{\partial t\,\partial y}\right| \approx \frac{d}{2f}\,l\omega\xi_0 \leq 2 \times 10^{-2}\xi_0$$

and

$$\left|\frac{1}{2g}u_0^2\right| \approx \frac{g}{2f^2}\,l^2\xi_0^2 \leq \frac{1}{2} \times 10^{-3}\xi_0$$

in which u_0 is given by $-g/f(\partial/\partial y)\xi_0$. In the last relation we assume further that $\xi_0 \leq 10^2$. In a lower layer the same estimation is valid.

Therefore we can neglect these inertia terms within an error of 2 pct. In other words, we have

$$f\bar{v} = \frac{g}{d}(\xi_0 - \xi_b) \tag{7.1'}$$

instead of (7.1). This equation indicates that the geostrophic relation is valid even in the western zone if we consider the average condition over the width.

Since (7.1') and (7.2) are linear, we can apply the method of separation into normal modes to obtain $\bar{v}$ and $\bar{v}'$. Then (7.1') and (7.2) become

$$dfV_i = g(R_{i0} - R_{ib}) \tag{7.4}$$

$$\frac{\partial}{\partial y} V_i - \frac{g}{fd}\frac{\partial}{\partial y} R_{i0} = -\frac{\kappa_j^2}{2H}\gamma_1 \frac{\partial}{\partial t}(R_{i0} + R_{ib}) \quad (j = r, l) \tag{7.5}$$

by use of (7.3), in which

$$V_i = \sigma_i \bar{v} + \overline{v'} \tag{7.6}$$

The elimination of R_{ib} in (7.4) and (7.5) gives the equation of $\bar{V}$,

$$\frac{\partial V_i}{\partial y} - C_i^{-1}\frac{\partial V_i}{\partial t} = \frac{g}{fd}\left(\frac{\partial R_{i0}}{\partial y} - 2C_i^{-1}\frac{\partial R_{i0}}{\partial t}\right) \tag{7.7}$$

with

$$C_i = 2gH/\gamma_1 \kappa_i^2 fd \tag{7.8}$$

In this equation the second term of both sides represents the effect of a divergence in the western zone. As before, we put $\partial/\partial y = l$; $\partial/\partial t = \omega$, then the ratio of the second term to the first one is given by

$$\left|C_i^{-1}\frac{\partial V_i}{\partial t}\right| \Big/ \left|\frac{\partial V_i}{\partial y}\right| \approx \frac{1}{3} \times 10^4 \omega \kappa_i^2$$

for the right hand side. This ratio is 5×10^{-3} in a barotropic mode and 19 in a baroclinic one. Therefore, in a barotropic mode the effect of a divergence in the western zone can be neglected within an error of 1 pct.

The equation (7.7) gives the waves propagating to the south with a phase velocity C_i. When we put

$$R_{i0} = P_i \cos(\omega t - \theta_i) \sin ly \tag{7.9}$$

following (6.13) and further assume the boundary condition

$$V_i = 0 \quad \text{at} \quad y = 0 \tag{7.10}$$

then the solution of (7.7) is given by

$$V_i = \frac{g}{fd} P_i \{(1 - r_{1i}) \cos(\omega t - \theta_i) \sin ly - r_{2i} \sin(\omega t - \theta_i) \cos ly + r_{2i} \sin(\omega t + C_i^{-1}\omega y - \theta_i)\} \tag{7.11}$$

in which

$$r_{1i} = \omega^2/(C_i^2 l^2 - \omega^2)$$
$$r_{2i} = C_i l\omega/(C_i^2 l^2 - \omega^2) \tag{7.12}$$

When we neglect the divergence in the western zone, the coefficients r_{1i} and r_{2i} in the equation (7.11) should vanish. Therefore these coefficients represent the effect of divergence. In the left hand side of (7.11) the last term gives the waves moving to the south. For a barotropic mode, $C_r = 0.8 \times 10^6$ and in a range of frequency considered here these coefficients are approximately given by

$$r_{1r} \approx \omega^2/C_r^2 l^2 = 0.83 \times 10^{-4} \times T^{-2}$$
$$r_{2r} \approx \omega/C_r l = 0.91 \times 10^{-2} \times T^{-1}$$

by taking the period T in days. Therefore they are negligibly small, as expected from the estimation of the terms in (7.7).

On the other hand, in a baroclinic mode $C_l = 51.1$ and at about $T = 143$ days these coefficients become infinite. This resonance occurs when the wave length is twice the north-south distance of the ocean. The curves of r_{1l} and r_{2l} are shown as functions of T in Figure 8. The response of a baroclinic mode near this resonance period belongs to the *SC* stage as seen from Figures 6 and 7, and the amplitude P_l itself in (7.11) is small compared with the equilibrium value. However, this period is very close to the predominant period of the fluctuation in the western current such as the Gulf Stream and the Kuroshio [*Iselin*, 1940; *Ichiye*, 1958] which show a semi-annual variation notwithstanding the annual change of the wind stresses. Further, near this resonance period the phase lag of the mass transport is not uniform along the western zone owing to the waves expressed in the last term of (7.11).

8. *Summary of the theory and application to a real ocean*—The response of the western current in an ocean to variable wind stresses is treated theoretically in this paper. The ocean is divided into the narrow western zone and wide interior region; in the former only the mass transport is considered, while in the latter the geostrophic approximation is applied. Then we can obtain the relation between the mass transport of the western current and the wind stresses both for a homogeneous or a two-layered ocean by use of linear equations only. The results obtained from this analysis are as follows:

(1) In the stationary state the mass transport derived from this method without a friction approaches the value computed from the solution of the frictional boundary model of *Stommel* [1948] within an error of 10 pct for a frictional coefficient k less than 7×10^{-7}.

(2) For periodic wind stresses there exist two types of Rossby waves in the interior region. Then there are two alternate conditions to determine the western mass transport. One is to eliminate one type of Rossby waves of a shorter wave length. Another is to make the currents zonal at the western boundary of an interior region. Both conditions lead to the same equilibrium state as the frequency of the wind stresses decreases.

(3) In a homogeneous ocean or in a barotropic mode of a two-layered ocean, the response approaches the equilibrium state for a period longer than a month with the phase lag of about three days behind the wind stresses.

(4) In a baroclinic mode the equilibrium state is attained for a period longer than 10^4 days with the phase lag of about 1500 days.

(5) There is a critical period of about 10 days in a barotropic mode and of about 210 days in a baroclinic mode. The amplitude of the response becomes very small for a period shorter than this, because the Rossby waves in this range are highly dispersive.

(6) In a baroclinic mode, the amplitude shows the spectral distribution between the critical period and the equilibrium stage. Thus the amplitude has peaks and nodes at certain periods.

(7) The effect of the inertia terms in the western zone is proved to be negligible when we consider the mass transport instead of the current itself.

(8) A divergence in the western zone has almost no effect on a barotropic mode. On the other hand, it produces the waves moving southward, which indicate the resonance at the period of 143 days in a baroclinic mode.

In this paper it might be wise to refrain from applying the theory set forth to a practical problem for two reasons. One is that observed data concerning the fluctuation of the transport of the western current have been too scarce and also too inaccurate to be compared with the theory. Another is that the simplest model is used here, both concerning wind stress distribution and geometry of an ocean, so as to understand a basic process underlying the mechanism of coupling between the wind system and western current rather than to apply the theory in explaining existent data.

However, if we examine observed facts carefully, we can detect the applicability of the theory even to very complicated phenomena like the variation of ocean currents. As one example, we should refer to the data about the mass transport of the Florida Current measured by *Wertheim* [1954]. In Figure 9 one of his results

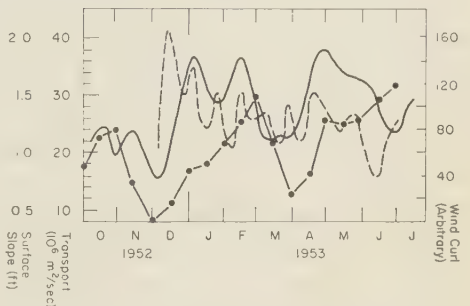


FIG. 9—The mass transport of the Florida Current between Key West and Havana (full line), sea surface slope between Bermuda and Charleston (broken line) and wind curl over the Atlantic (dot and line) after Wertheim [1954]

is given in which the mass transport measured by a submarine cable across the Florida Straits, the difference of tide gauge level between Bermuda and Charleston which is considered as a measure of the Gulf Stream transport, and wind torque over the Atlantic Ocean are plotted. It is rather disappointing that these two kinds of transport indicate no correspondence to the wind torque, as predicted by the theory. An

excuse may be found in a switching mechanism for a discrepancy between the Florida Current transport and the wind torque, as pointed out by Wertheim.

However, there seems to be no reason for a discrepancy between the sea level difference data and the wind torque, since in the last section we have proved that the geostrophic relation is valid even in the western zone for the mass transport. On the other hand, the fluctuation of the wind torque in Figure 9 indicates a period of about four to five months from October, 1952, to April, 1953. This period seems to be responsible for causing the resonance phenomenon in the baroclinic mode and thus exciting a large fluctuation in the mass transports of the Florida Current and the Gulf Stream. At the same time, the simultaneous correlation between the mass transports and wind torque may be disturbed by such a resonance as shown in the equation (7.15). Further, *Wertheim* [1954] computed the correlograms of wind stress and transport of the Florida Current for a longer period than is shown in Figure 9 and found a positive correlation of about 0.62 with $a - .5$ month phase lag of transport behind the wind. He rejected this value as fortuitous, but it seems possible that this suggests the simultaneous change of wind stress and western mass transport for other periods without resonance, because the predicted value of about three days phase lag by the theory is undetectable in the observation.

Anyhow, the process of the present analysis might be very useful for understanding the mechanism of the change of the western current, since linearized equations are used in a whole ocean, including the western current zone. The fundamental reason for a validity of the separation of the ocean into two parts, even in transient stage, has been stated in *Ichiye* [1958]. The energy applied to the ocean by wind stress in the eastern portion propagates westwardly and thus influences the western current, but the energy given in the western portion is damped out quickly and has no effect on the eastern portion. This situation is also explained from the general theory of the Rossby waves in the ocean [*Arons* and *Stommel*, 1956]. The Rossby waves of a longer wave length have a westward group velocity and serve to transport the energy given to the ocean by a large scale wind system towards the west. On the other hand, the Rossby waves of a shorter wave length have an eastward group velocity; but the wave length becomes so short for a period of a week or longer that it is attenuated as it moves eastwardly by a slight dissipative force which is inversely proportional to the wave length. As a result, the interior region has an influence on the western zone, but the reverse is not true.

In the present model we use the simplest distribution of wind stress distribution for the stated reasons; however it is not difficult to develop the theory to a more general kind of wind stress distribution, except in computational complications. The development of the theory in this direction, to say nothing of an assumption of a more realistic geometry of an ocean, may help much in the applicability of the theory to practical purposes. This can be done either by assuming wind stresses of any arbitrary function of time or by giving them an arbitrary spatial distribution. Mathematically this corresponds to giving τ_0 (amplitude of wind stress) of the equation (4.3) as a function of ω or p. Then we can obtain a general solution by integrating the Fourier integral of the solution of R_i about ω or p. Particularly in such an ocean as the Pacific where the wind stress distribution is not uniform zonally, the results concerning the change of the western mass transport are much affected by the zonal structure of the wind and the present theory becomes less applicable.

Acknowledgments—This study was aided by a contract between the Office of Naval Research Department of the Navy, and the Florida State University, NR 163-396. The author wishes to thank Henry Stommel, who initiated his interest in this problem; F. C. W. Olson, who gave helpful advice; Sidney W. Fox, who encouraged him throughout the present work; and Horace Loftin who gave assistance in preparing the manuscript. This paper is Contribution No. 94 of the Oceanographic Institute, Florida State University.

References

Arons, A. B., and H. Stommel, A β-plane analysis of free periods of the second class in meridional and zonal oceans, *Deep Sea Res.*, **4**, 23–31, 1956.

Charney, J. G., The Gulf Stream System as an inertial boundary layer, *Proc. Nat. Acad. Sci.* (USA), **41**, 731–740, 1955a.

Charney, J. G., The generation of ocean currents by wind, *Jour. Mar. Res.* **14**, 477–498, 1955b.

Ichiye, T., On the variation of ocean circulation (1), *Oceanog. Mag.* **3**, 79–82, 1951.

Ichiye, T., The response of a stratified, bounded ocean to variable wind stresses, *Oceanog. Mag.* **10**, Woods Hole Oceanographic Institution Contr. No. 963 (in press), 1958.

Iselin, C. O'D., Preliminary report of long period variations in the transport of the Gulf Stream, *Pap. Phys. Oceanog. Met.*, **3**, 1–40, 1940.

Morgan, G. W., On wind-driven ocean circulation, *Tellus*, **8**, 301–320, 1956.

Munk, W. H., On the wind-driven ocean circulation, *J. Meteorol.*, **7**, 79–93, 1950.

Munk, W. H., Note on Charney's calculation, (Unpublished manuscript), 1957.

Stommel, H., The westward intensification of wind-driven ocean circulation, *Trans. Amer. Geophys. Union*, **29**, 202–207, 1948.

Stommel, H., Florida Straits transports, 1952–1956, *Bull. of Marine Sci. of the Gulf and Caribb.*, **7**, 252–254, 1957.

Veronis, G., and G. W. Morgan, A study of the time-dependent wind-driven ocean circulation, *Tellus*, **7**, 232–242, 1955.

Veronis, G., and H. Stommel, The action of variable wind stresses on a stratified ocean, *Jour. Mar. Res.*, **15**, 43–75, 1956.

Von Arx, W. S., D. F. Bumpus, and W. S. Richardson, On the fine structure of the Gulf Stream Front, *Deep Sea Res.*, **3**, 46–65, 1955.

Wertheim, G. K., Studies of the electric potential between Key West, Florida and Havana, Cuba, *Trans. Amer. Geophys. Union*, **35**, 872–882, 1954.

(Manuscript received October 20, 1958.)

The Great Lakes Storm Surge of May 5, 1952*

William L. Donn

Brooklyn College and
Lamont Geological Observatory
(Columbia University)
Palisades, New York

Abstract—A damaging storm surge struck much of the shore zones of Lakes Huron and Erie on the morning of May 5, 1952. The surge, which was associated with a moving atmospheric pressure-jump of average characteristics, is explained on the basis of resonant coupling to both edge waves and gravity waves. The latter in turn created above-normal seiches in lake areas with appropriate dimensions and forms. A pronounced lag occurred between the surge and the generating air disturbance at places where lake parameters did not permit resonant coupling, with the surge arriving from regions where coupling was possible.

Introduction

Earlier empirical and theoretical studies of certain storm surges in Lake Michigan [*Ewing* and others, 1954; *Donn* and *Ewing*, 1956] have explained these on the basis of resonant coupling between either gravity waves or edge waves and a fast moving atmospheric pressure jump line with associated high winds. The cases studied were confined to the shores of the southern half of Lake Michigan. The surge described below involves several of the Great Lakes and was first reported in newspapers describing heavy damage from 'freak' tidal waves that battered the Canadian and American shores of Lake Huron, and the shores of the St. Clair River and Lake Michigan in the vicinity of Grand Traverse Bay. Correspondence from official weather observers at Sarnie, Ontario, on the west shore of Lake Huron, describes lake waves that were phenomenal while the related atmospheric disturbance, although definite, was not unusual in intensity. Also, limnograms (lake level records) from Cleveland and Buffalo on Lake Erie showed high waves which continued for some time after the first arrivals.

This study is based on available data from the limnograms made at Harbor Beach and Mackinaw City on Lake Huron, and Cleveland and Buffalo on Lake Erie (see Fig. 1 for locations). Most of Lake Michigan was not affected by this surge, and although newspaper accounts indicate that the surge was quite widespread over the Great Lakes, no adequate data were available from other shore and lake areas.

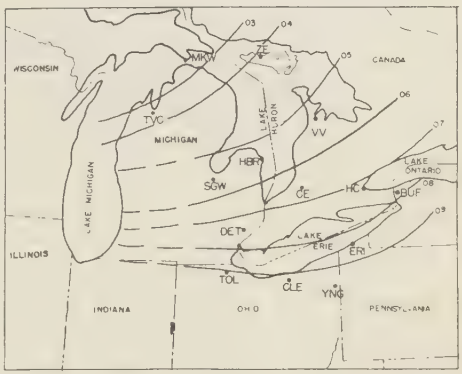


Fig. 1—Map of Great Lakes region showing isochrones for the pressure jump line of May 5, 1952 (Station symbols are explained in Table 1)

The Data

Pressure data—Examination of the synoptic meteorological data about the time of the lake surges indicates that a strong pressure disturbance had crossed the critical area on the morning of May 5. Data obtained for a number of stations (located in Fig. 1) show the occurrence of a pressure jump whose arrival times and magnitude at the stations are summarized in Table 1.

Although the pressure change showed local variation, its value averaged about 0.1 inches. Based on this data from fourteen stations, isochrones for a traveling pressure jump line have been constructed as in Figure 1. It is

* Lamont Geological Observatory Contribution No. 326.

TABLE 1

City	Pressure jump arrival time EST	Pressure jump magnitude
Buffalo, N. Y. (BUF)	0800	.12 inches
Centralia, Ont. (CE)	0630	.15
Cleveland, Ohio (CLE)	0928	.08
Detroit, Mich. (DET)	0738	.07
Erie, Pa. (ERI)	0800	.10
Harbor Beach, Mich. (HBR)	0520	*
Macknaw City, Mich. (MKW)	0236	*
Saginaw, Mich. (SGW)	0540	.06
Toledo, Ohio (TOL)	0940	.02
Traverse City, Mich. (TVC)	0345	.02
Youngstown, Ohio (YNG)	0930	.10
Gore Bay, Ontario (ZE)	0430	.10
Hamilton, Ontario (HC)	0730	*
Wiarton, Ontario (VV)	0515	.14

* Only arrival time available.

TABLE 2

City	Wind velocity
Buffalo, New York	On arrival of pressure jump wind shifts from weak east to weak north
Cleveland, Ohio	Shift from northeast to southeast at 20 to 30 mph on arrival of pressure jump
Detroit, Michigan	On arrival of pressure jump wind shifts from southeast to northeast; 25 to 39 mph
Erie, Pennsylvania	22 mph from the north; on arrival of jump wind shifts to southeast
Mackinaw City, Michigan	36 mph with wind shifts at 0236
South Bend, Indiana	18 mph from northwest; on arrival of jump wind shifts northeast to north-northeast
Toledo, Ohio	On arrival of jump wind shifts from south to north 12 to 15 mph
Traverse City, Michigan	On arrival of jump wind shifts from southeast to east 20 to 25 mph
Youngstown, Ohio	On arrival of jump wind shifts from north to southeast 30 to 35 mph
Centralia, Ontario	On arrival of jump wind shifts from east-southwest to north-northeast
Gore Bay, Ontario	Wind shifts from east to north; 10 to 26 mph
Hamilton, Ontario	Wind from northeast at 10 mph
Wiarton, Ontario	Increase from 3 mph to 17 mph with wind shift from northeast to north-northeast

evident that the disturbance moved roughly northwest to southeast with both variable speed and direction. The pressure jump moved with the highest speed, 65 mph, from 0400 to 0500 when it traversed the central area of Lake Huron. When crossing Lake Erie, between 0700 and 0900, it slowed to an average speed of 32 mph. It is of great importance to note that the disturbance line appeared to travel transverse to the shores of Lake Huron and parallel to the shores of Lake Erie. Average speed across the entire critical area was 36 mph.

Wind data—Recorded wind speeds accompanying the pressure jump line were variable, being quite high at some stations and small at others. For the most part the wind shift toward the north or northeast occurred at about the same time as the pressure jump. This provided a check on the times of the pressure change. A summary of the wind observations is given in Table 2.

Water level data—Limnograms have been studied from four stations located in Figure 1 as MKW (Mackinaw City), HBR (Harbor Beach) CLE (Cleveland) and BUF (Buffalo).

Tracings of each of these records are reproduced in Figure 2. The tracing of the Mackinaw City record shows the first and relatively weak disturbance beginning just after 0200, May 5, at the time of pressure jump passage. A stronger surge consisting of waves one half foot in height and about 90 min period began 7.5 hours subsequent to the first event and continued for about 10.5 hours. (Note that the record reads from right to left.)

According to the tracing the main surge at Harbor Beach began at 0520, coinciding with the arrival of pressure jump, and continued for

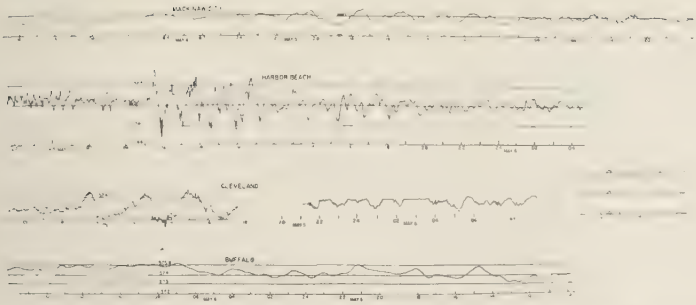


FIG. 2—Limnograms (water level records) showing the storm surge at four stations during May 5, 1952 (Mackinaw City and Buffalo are read from right to left)

more than six hours with an average period of 67 min and height of 2.2 ft. Wave delineation becomes somewhat irregular about 0900. This will be referred to below in the discussion. Beginning about 1200, a secondary set of waves of lower height but longer period (84 min) becomes evident. These waves continue for several hours with slow attenuation. Superimposed on all of the long period waves are the interference effects of fairly strong, shorter period background waves which are prominent on the record prior to the surge.

The main surge at Cleveland (Fig. 2) begins at 0900 with a duration of at least 10 hours. The average period for these waves is about 180 min, and the height 1.7 ft. Following the record interruption, the waves continue in an attenuated and somewhat irregular form into May 6.

The Buffalo limnogram tracing is also shown in Figure 2 which again reads from right to left. This shows an amplitude increase just after 1300 on May 5, about four to five hours after the arrival of the surge at Cleveland. The principal waves which have an average period of 3.3 hrs and a height close to one foot can be detected for at least 16 hrs. Note that the vertical scale reduction here compared with the three previous records gives a misleading picture of height at first glance.

DISCUSSION

Although the disturbance was experienced first at Mackinaw City, it will be more convenient to discuss this aspect after consideration of the surge at Harbor Beach.

Harbor Beach—The relation of Harbor Beach to the coastline and bottom topography of Lake Huron is shown in Figure 3. The central area of the lake bottom is quite irregular owing to the drowned pre-glacial stream valleys. However, the bottom slope along the coast from Thunder Bay southward and around to the Saugeen Peninsula is fairly uniform. According to Figure 1 the pressure jump line was oriented transverse to the shore as it traveled southward across the long dimension of the lake.

The combination of uniform off-shore slope and trends of the shore line and pressure jump line seem to be quite conducive to edge wave rather than gravity wave generation as an explanation for the Harbor Beach surge. The shore zone between Thunder Bay and Harbor Beach would be the likely fetch area for edge waves arriving at Harbor Beach at the time of the atmospheric pressure jump. Edge wave velocity is given by the formula

$$c = \frac{Tg \sin \beta}{2\pi}$$

where T is a wave period, β is the slope of the bottom, and g, the acceleration of gravity. Using the observed wave period, the velocity is calculated for profiles 1 to 4 (Fig. 3) as 58, 43, 45, and 49 mph, with an average velocity of 49 mph. An average edge wave velocity of 45 mph exists for profiles 5 and 6, between Saginaw Bay and Harbor Beach. According to Figure 1 the pressure jump line traveled southward at 50 to 60 mph along the shore between the profiles 1 to 4 but slowed to about 38 mph along profiles 5 to 6. Calculated edge wave velocities thus lie within the velocity limits of the pressure jump line.

Further, as the pressure jump line is in reality

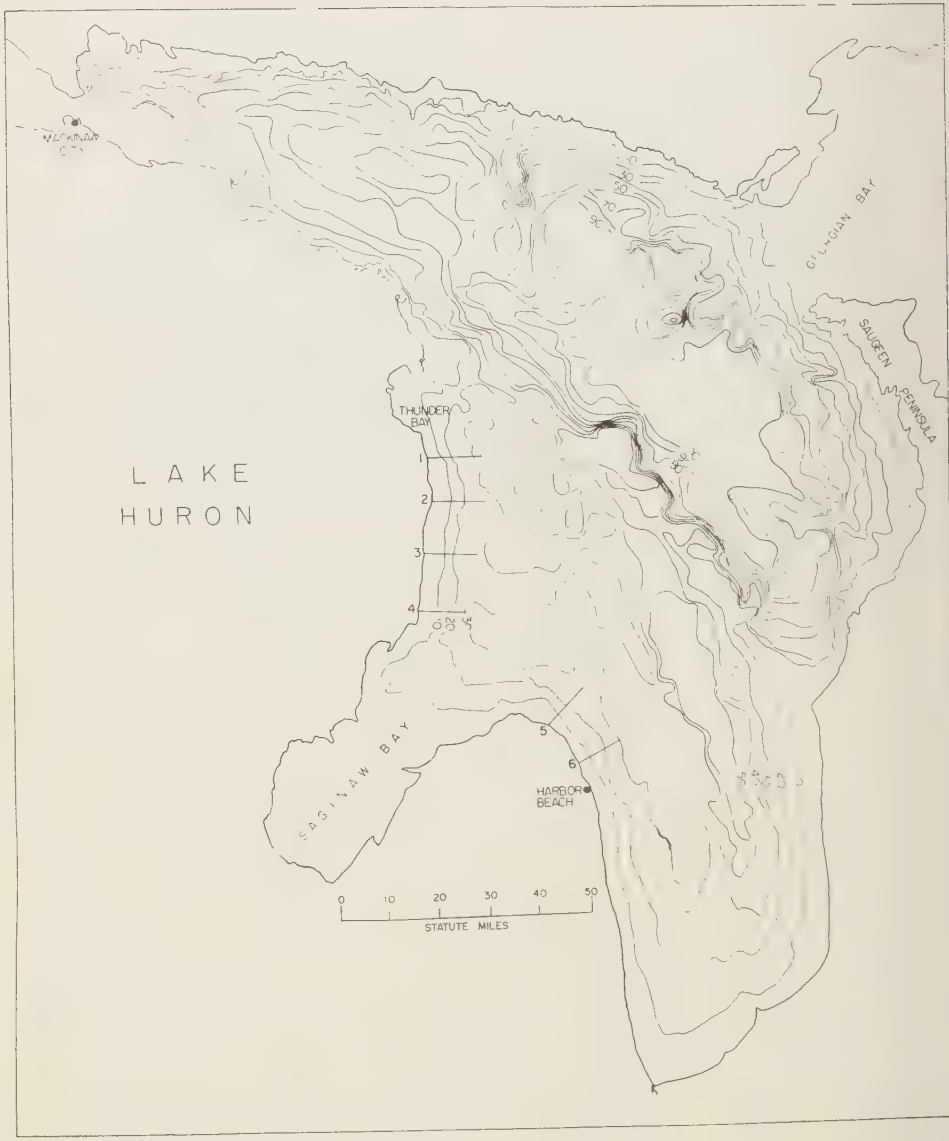


FIG. 3—Bottom topography of Lake Huron

a band of some width, conditions for resonant transfer of energy between the air disturbance and the lake waves seem good. As the wave length is large compared with the entrance width of Saginaw Bay, the effect of the bay in seriously modifying the waves is considered negligible. Conditions along the shore of Lake Huron are thus very similar to those in Lake Michigan as described by *Donn* and *Ewing* [1956] when edge wave theory was first applied to the Great Lakes. More complete discussions of edge wave theory have been given recently by *Ursell* [1951, 1952]; *Munk* and others [1956]; and *Donn* and *Ewing* [1956].

The observed duration of the prominent disturbance at Harbor Beach is about 6.5 hrs. A theoretical duration can be determined based on the size of fetch and the wave velocities. Hence, using profile 1 (Fig. 3) as the beginning of the fetch for the western side of the lake, the duration of the disturbance should be about two hours based on a phase velocity of 58 mph for the front of the water disturbance and a group velocity of 29 mph (c/2) for the rear of the disturbance. However, it is clear from Figures 1 and 3 that similar edge wave conditions are present along that eastern shore of the lake which begins just opposite to Harbor Beach. Reports indicate that similar surge activity was observed in this region.

Using an average speed of 56 mph (from the pressure jump isochrones) for the front of an edge wave disturbance generated along the eastern margin of the lake and beginning opposite to profile 1, the travel time to Harbor Beach would be 3.6 hrs and the disturbance would have a duration of close to four hours. The disturbance from the eastern side should arrive after 0840 and continue after 1220 giving a total duration of 6 hr 40 min at Harbor Beach, a value which coincides well with the observation. It can be noted in Figure 3 that the first two or three waves of the main surge are quite regular and easy to delineate, with the irregularity which begins at 0900 being explained by the interference beginning with the arrival of the wave disturbance from the eastern side of the lake. This agreement between theory and observation tends to confirm the edge wave nature of the surge.

According to theory, amplitudes of edge waves decrease exponentially from the shore, increasing further the difficult problems of amplitude estimation. Although the barometric equilibrium wave in the water would be only about one inch, *Lamb* [1932] showed that considerable amplitude increase would occur as velocities of the water and moving air disturbance approached each other. Further recent experiments by *Wiegel* and others [1958] show tank waves under conditions of resonance with amplitudes about 25 times that of equilibrium waves. The effect of the variable but often high winds over the lake will further increase amplitudes.

Following the main surge at Harbor Beach, a prolonged and slowly attenuating 'coda' was noted above, with a wave period of about 85 min. As in the case of the Lake Michigan surge described by *Donn* and *Ewing* [1956] this effect seems explainable by the excitation of a transverse seiche in the southern part of the lake from the passage of the severe atmospheric disturbance. The calculated seiche period for an average transverse profile near Harbor Beach is 95 min, which agrees well with the 85-min observed waves, the difference being within the experimental error of record reading and profile averaging.

Mackinaw City—Because Mackinaw City is located on the constricted Straits of Mackinac connecting the narrow northern portions of Lakes Huron and Michigan, there is little fetch for the development of a significant direct water disturbance. According to Figure 1, the atmospheric pressure surge crossed the station just prior to 0300. At this time a weak disturbance is distinctly visible on the limnogram and probably marks the direct pressure and wind effect on local water level.

However, a low amplitude, long period wave group begins a few hours later and reaches maximum amplitude of 0.5 ft some 15 hrs later. These waves are probably a combination of edge waves traveling northward from around the southern edge of the lake and possible oscillations of the lake in the northwest-southeast basin shown by the contours. These latter waves may be likened to a reverberation effect from the disturbance. Although none of the later events is distinct enough to warrant precise application of theory, this case certainly indicates, when compared with Harbor Beach, the importance of resonant coupling over a significant fetch, for the later arrivals generated by such coupling at a distance are much stronger than the direct arrival. This is identical with a similar situation at Ludington on Lake Michigan described earlier by *Donn* and *Ewing* [1956].

Cleveland—The atmospheric disturbance reached Cleveland in the form of a hard squall and showed a pressure jump of nearly 0.1 inch and winds from 20 to 30 mph (recorded at 0928 about ten miles in from shore). The surges in air and water must therefore have reached the recorder simultaneously. It can be estimated from Figure 1 that the pressure jump, which traveled as a line or band oriented parallel to the long dimension of Lake Erie, slowed to about

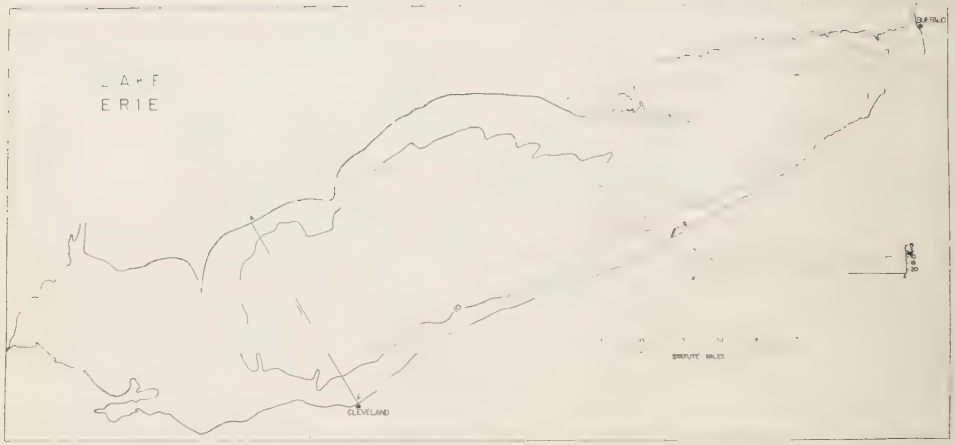


Fig. 4—Bottom topography of Lake Erie

27 mph while crossing the lake. It is evident from the map and profile in Figure 4 that the bottom of Lake Erie is quite flat. Based on the depth along Profile A-A′, the velocity of a long gravity wave here is about 30 mph, making it quite reasonable that resonant transfer of energy occurred between the air disturbance and a long gravity wave generated in the lake. The relatively high surge that arrived simultaneously with the pressure jump and squall can thus be explained by resonant coupling to a gravity wave, as was noted above for the edge waves at Harbor Beach.

The first set of high amplitude waves with an average period of 180 min is followed after the record break by an attenuated irregular train of waves with prolonged duration. Owing to instrumental difficulties, the exact duration of the large initial waves cannot be determined, but is at least 10 hr. It is clear then that this prolonged train could not be the direct effect of the incident water disturbance which would be less than two hours if the front is considered to travel with the phase velocity (30 mph) and the rear with the group velocity (15 mph).

This wave train might be considered from the viewpoint of repeated reflections. Using the phase velocity, the time for one complete reflection is about 3.4 hr, which agrees well with the first and third crest intervals which are 3.2 and 3.4 hr, and somewhat less well with the second interval which is about 2.5 hr. However, lack of any attenuation for the first few waves tends to weaken the argument for repeated reflection.

A third explanation for the prolonged train is a seiche set up by the initial water transfer to the southern end of the lake. Using the formula for the fundamental mode in a closed rectangular basin, $T = 2L/\sqrt{gh}$, where T is the period, L the length, g the acceleration of gravity, and h the depth, the seiche period for the Profile A-A′ is about 197 min. This agrees well with the observed average period of 180 min with the difference being within the experimental error, as there is some subjectivity in determining the profile and depths to be used, and as the formula is not strictly applicable to a lake with the form of Lake Erie. The prolonged attenuated waves evident after the record interruption are probably a continuation of attenuated seiche plus a reverberation effect from around the lake.

Buffalo—Located at the narrow eastern end of Lake Erie, Buffalo experienced the pressure jump at 0800, yet the first water disturbance began five to six hours later, between 1300 and 1400, according to the limnogram in Figure 3. (Note that the record reads from right to left.) Prior to 1200 the trace was essentially at background level. The principal Buffalo oscillations, which continued into the early part of the following day, were almost as high as those in Cleveland.

It seems clear from the geography of eastern Lake Erie that little fetch, with or without possible resonance, could be expected for the pressure jump in the vicinity of Buffalo, hence the absence of simultaneous air and lake disturbances.

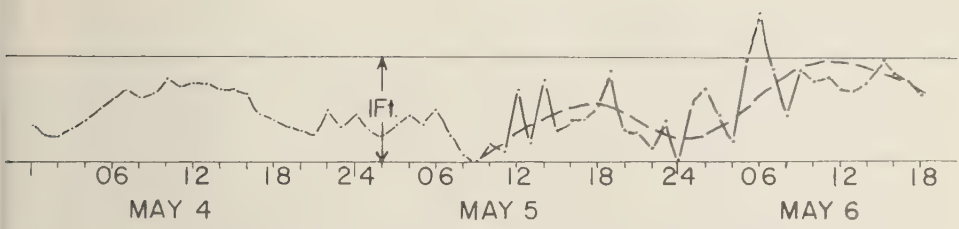


FIG. 5—Buffalo limnogram compressed to show long-period seiche

It is difficult to explain the relatively high, long-period waves commencing at about 1400 in terms of a seiche for several reasons. The reason for the five to six hour lag in development is not evident if the waves are of seiche origin. Further, if the seiche were an oscillation in the transverse or northwest-southeast direction, Buffalo would probably be at or near a nodal line and thus experience little vertical water change. The theoretical seiche period for the longer dimension is about 16 hr, or much too long for that observed, and even the seiche period for the smaller partially enclosed northeastern basin is about 6 hr if the basin be considered completely enclosed, and 12 hr if the basin be treated more properly as being opened at one end.

A more satisfactory explanation seems to be one of propagation from a more distant area of the lake. The oscillations already discussed for Cleveland are of nearly identical periods as those recorded at Buffalo, where the first main crests arrived five hours later. Edge wave propagation from the western portion of the lake in the vicinity of Cleveland seems to provide a possible explanation for the later Buffalo disturbance. In the previously published studies of edge waves in Lake Michigan, Donn and Ewing found wave lengths to be of the order of 80 miles. As the edge wave effect is limited to a distance $\lambda/2\pi$, the offshore slope taken to a distance of 12 miles out from Cleveland is about 0.001. Using this value in the velocity formula for edge waves given earlier, a value of about 45 mph is obtained for edge wave velocity for the wave period observed at Buffalo.

An edge wave disturbance generated near Cleveland would thus arrive at Buffalo about four hours later. As the factors of wave length and bottom slope and point of wave origin involve estimates, the wave travel time can be revised upward or downward, depending upon values used, but reasonable estimates seem to give good results in explaining the Buffalo waves. Amplitudes of the larger waves at Buffalo would be reinforced by the pronounced funneling and shoaling effects at the western margin of Lake Erie.

In an attempt to observe any long-period seiche related to oscillations of the entire lake, the Buffalo data were replotted on a compressed time scale from hourly measurements. An oscillation of very close to the theoretical period can be seen on the record of May 4, and it shows a somewhat increased amplitude through the storm surge waves on May 5 and 6 (Fig. 5). A separate and more complete study of these oscillations now under way indicates that this period occurs frequently from excitation by storm surges and other stimuli.

CONCLUSIONS

This study involves a variety of hydrodynamic oscillations generated by the same disturbance under different geographic conditions. It confirms for Lake Huron and Lake Erie the existence of the same kinds of lake disturbances already described for Lake Michigan. Maximum effects occur when resonant transfer of energy is possible between the moving disturbance and the water body in question. For the most part, either long gravity waves or Stokes' edge waves are generated, depending on the relation between the orientation of the air disturbance and local lake geography. Very prominent seiches are excited when the edge wave (or gravity wave) period is close to the natural period of oscillation of the parts of the lake being disturbed. It now seems clear that, depending on speed and path, an atmospheric disturbance of moderate intensity can vary from a harmless to a dangerous situation when over the Great Lakes or a similar water

body. Further experimentation coupled to theoretical deductions from existing gravity wave, edge wave, and seiche formulas should make it possible to predict resonant velocities for these disturbances for all parts of all of the lakes and provide a good warning system for future lake surges generated by moving disturbances in the atmosphere.

Acknowledgments—This study was supported in large part by a grant from the Engineering Foundation in New York City. Limnograms from all of the Great Lakes were furnished by the United States Lake Survey in Detroit, Michigan, and necessary weather data were supplied by numerous local offices of the United States Weather Bureau, in addition to the National Weather Records Center in Asheville, North Carolina. Miss Anita Fishman was of considerable aid in processing data and in preparing the report. The writer is very grateful to all of these individuals and organizations who made this study possible.

References

Donn, W. L., and M. Ewing, Stokes' edge waves in Lake Michigan, *Science*, **124**, 1238–1242, 1956.

Ewing, M., F. Press, and W. L. Donn, An explanation of the Lake Michigan Surge of 26 June, 1954, *Science*, **120**, 684–686, 1954.

Lamb, H., *Hydrodynamics*, Cambridge Univ. Press, Cambridge, 1932.

Munk, W., F. Snodgrass, and G. Carrier, Edge waves on the continental shelf, *Science*, **123**, 127–132, 1956.

Ursell, F., Trapping modes in the theory of surface waves, *Proc. Cambridge Phil. Soc.*, **47**, 3, 1951.

Ursell, F., Edge waves on a sloping beach, *Proc. Roy. Soc. London*, **A214**, 79, 1952.

Wiegel, R. L., C. M. Snyder, and J. E. Williams, Water gravity waves generated by a moving low pressure area, *Trans. Am. Geophys. Union*, **39**, 224-236, 1958.

(Manuscript received October 23, 1958.)

Wave Forces on Groups of Vertical Cylinders*

J. E. Chappelear

Shell Development Company
Houston, Texas

Abstract—A method is presented for the calculation of wave forces on groups of circular cylinders. The assumptions employed are that (1) wave forces are produced by a localized nonlinear phenomenon, for example, a detached boundary layer, and the velocity distribution around the pilings is that given by potential flow; (2) the center-to-center distances of the various cylinders are small compared with the wavelength (Airy theory); (3) the radii are small compared with the center-to-center distances; (4) the wave force on one cylinder is known empirically. Then a calculation of the laminar velocity field around the cylinders can be made. The result obtained is that only the magnitude of the velocity distribution and the location of the stagnation points are changed, not the details of the angular dependence. The angular dependence is the same as is calculated on one cylinder. Then, arguing that similar causes produce similar effects, we can calculate geometrical correction factors to the drag and inertial coefficients for a single cylinder. Certain examples involving a small number of cylinders in simple geometrical arrays are presented to illustrate the application of the method, and a brief discussion of the experimental evidence is given.

I. Introduction

Little theoretical consideration has been given to the problem of surface wave forces on complex structures. Even in the case of small waves on a single cylinder the calculations are so complicated and the theory in such an unreliable state that no conclusions could be drawn; hence, it was considered useless to attempt a more complicated problem. Some progress can possibly be made for simple geometrical configurations of structural elements. In particular, groups of circular cylinders seem to be the most promising cases to investigate.

It is found that for circular cylinders the problem can be solved phenomenologically. That is, we can correct the forces on single pilings, which can be assumed to be known from experiments, by simple geometrical factors. It is not readily apparent that this particularly simple feature of the calculation can be extended to cylinders of irregular shapes, but this question was not pursued to any definite conclusions.

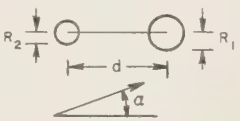
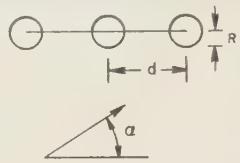
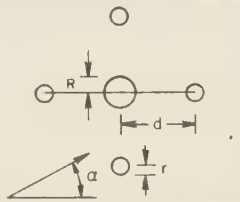
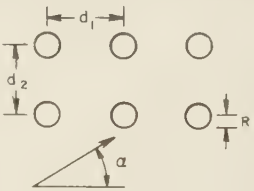
We do not consider the possible effects of vortex shedding as a mechanism for the interaction of neighboring cylinders.

For a steady flow past a single cylinder the presence of a wake behind the cylinder is sufficient to obtain a qualitative understanding of the force which is observed. More extensive experiments are necessary to understand wave forces, and until these experiments are made it is reasonable to assume that the mechanism of production of the drag force is nearly the same as in steady state. The fact that the formula of *Morison* and others [1954], which is a simple generalization of the steady-state engineering formula for wind forces on a cylinder, is useful in the correlation of published wave-force data is a good indication that there is considerable similarity between the steady and unsteady cases.

An experimental study has been made of the effects of such interference for the case of three vertical pilings [*Morison* and others, 1954]. The results were in general accord with the calculations of this report, but the inclusion of rather steep waves, including breakers, renders a direct comparison of magnitudes of the effects meaningless. The general result was that the moment on the center piling in the group of three was enhanced when the line of centers was perpendicular to the direction of wave advance, and diminished when the line of centers was parallel to the wave crests. This result is not unexpected in view of the results presented in Table 1.

* Publication No. 180, Shell Development Company, Exploration and Production Research Division, Houston, Texas.

TABLE 1

2 cylinders of radii R_1 and R_2 with separation d. α is the angle between the wave direction and the axis.		On R_1, $1 - \left(\frac{R_2}{d}\right)^2 \cos 2\alpha.$ On R_2, $1 - \left(\frac{R_1}{d}\right)^2 \cos 2\alpha.$
3 cylinders of radius R with separation d. α is the angle between the wave direction and the line of centers.		On left, $1 - \frac{5}{4}\left(\frac{R}{d}\right)^2 \cos 2\alpha.$ On center, $1 - 2\left(\frac{R}{d}\right)^2 \cos 2\alpha.$
4 cylinders of radius r whose centers are on a square of diagonal $2d$ and one cylinder of R at the center of the square. α is the angle between the wave direction and the diagonal.		On center, 1. On right, $1 - \left[\left(\frac{R}{d}\right)^2 + \left(\frac{r}{d}\right)^2\right] \cos 2\alpha.$
6 cylinders of radius R on a rectangular lattice $2d_1$ long and d_2 wide. α is the angle between the wave direction and a line through 3 cylinders.		On lower left, $1 - \left[\frac{2}{d_1^2}\left(\frac{R}{d_1}\right)^2 - \left(\frac{R}{d_2}\right)^2 + \frac{R^2(d_1^2 - d_2^2)}{(d_1^2 + d_2^2)^2} + \frac{8R^2(4d_1^2 - d_2^2)}{(4d_1^2 + d_2^2)^2}\right] \cos 2\alpha + \left[\frac{2d_1 d_2 R^2}{(d_1^2 + d_2^2)^2} + \frac{4d_1 d_2 R^2}{(4d_1^2 + d_2^2)^2}\right] \sin 2\alpha.$ On lower center, $1 - \left[\frac{2R^2}{d_1^2} - \frac{R^2}{d_2^2} + \frac{2R^2(d_1^2 - d_2^2)}{(d_1^2 + d_2^2)^2}\right] \cos 2\alpha.$

II. Statement of the problem*

We wish to calculate the force on a group of vertical, circular cylindrical objects subject to wave action. We can linearize the equations of hydrodynamics [*Lamb*, 1932] describing the motions of an inviscid fluid subject to conservative extraneous forces, and then by introducing a velocity potential, we obtain from the equations of motion an integral (Bernoulli's theorem) which enables us to calculate the pressure anywhere within the fluid and serves as a boundary condition on the free surface where the pressure is constant. This equation is

$$\frac{p}{\rho} + \Omega = \frac{\partial \tilde{\varphi}}{\partial t}. \tag{1}$$

From the continuity equation we obtain the differential equation for the velocity potential,

$$\nabla^2 \tilde{\varphi} = 0, \tag{2}$$

Laplace's equation. The boundary conditions are no flow through a rigid boundary (n is the normal to such a boundary),

$$n \cdot \nabla \tilde{\varphi} = 0, \tag{3}$$

and a radiation condition at infinity (no incoming scattered waves). In these equations $\tilde{\varphi}$ is the velocity potential ($u = -\nabla\tilde{\varphi}$), p is the pressure, ρ the density, and Ω the potential of the extraneous force (gravitation) per unit mass. The linearization is equivalent to an expansion in powers of H/λ (wave height divided by wavelength) and a retention of the first (linear) term of the velocity potential. On the free surface p/ρ is a constant. If we take the derivative with respect to time of equation (1), we obtain

$$\frac{\partial \Omega}{\partial t} = \frac{\partial^2 \tilde{\varphi}}{\partial t^2}, \tag{4}$$

on the free surface. Now Ω is gy, and its time derivative is the particle velocity on the free surface. To a linear approximation in H/λ, equation (4) is to be evaluated at mean water level ($y = D$), where to the same approximation the particle velocity is vertical. Thus equation (4) becomes, by means of the definition of the velocity potential,

$$-g \frac{\partial \tilde{\varphi}}{\partial y} = \frac{\partial^2 \tilde{\varphi}}{\partial t^2}. \tag{5}$$

*See 'List of symbols' at end of paper.

For a wave problem we can assume a harmonic time-dependence,

$$\tilde{\varphi} = Re\,(\Phi e^{-i\omega t}), \tag{6}$$

where ω is the angular frequency.

Thus we obtain from equation (5)

$$g \frac{\partial \Phi}{\partial z} = \omega^2 \Phi, \tag{7}$$

on $z = D$. To satisfy equation (7), and also to satisfy equation (3) on $z = 0$, it is convenient to put

$$\Phi = A\varphi(x, y) \cosh kz, \tag{8}$$

where A and k are constants at our disposal, and φ is an unknown function of x and y. We determine k to satisfy equation (7),

$$gk \sinh kD = \omega^2 \cosh kD. \tag{9}$$

As will be seen later we choose the one real root of this equation to obtain traveling waves. The substitution expressed by equation (8) is possible because we have vertical cylinders. Thus we can satisfy all the boundary conditions which contain z, and eliminate z from the problem.

To evaluate A we employ equation (1), noticing that far away from our scattering centers, we would like to have an incident plane wave.

$$\varphi = g \frac{H}{2} Re\, e^{i(kx - \omega t)}$$

$$= \frac{\partial}{\partial t} A \cosh kD\, Re\, e^{i(kx - \omega t)}. \tag{10}$$

We obtain for A that

$$A = -\frac{gH}{2i\omega \cosh kD}. \tag{11}$$

This same calculation for plane waves is given by *Lamb* [1932]. By means of these manipulations we have obtained a two-dimensional problem in x and y, which will now be restated. The differential equation to be solved is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k^2\right)\varphi = 0. \tag{12}$$

The boundary conditions are

$$n_x \frac{\partial \varphi}{\partial x} + n_y \frac{\partial \varphi}{\partial y} = 0, \tag{13}$$

where (n_x, n_y) is the unit normal vector to the various cylinders, and

$$\varphi \underset{\substack{x\to\infty \\ y\to\infty}}{\longrightarrow} e^{ik(x\cos\alpha + y\sin\alpha)} \tag{14}$$

+ outgoing scattered wave.

We see that k must be real for traveling waves.

III. 'Static' or long wavelength approximation

The idea of the approximation is simple. If our group of objects can be surrounded by a circle whose radius is small in comparison with a wavelength $(2\pi/k)$, then k, measured in reciprocal units of this radius, will be small in comparison with one. We can expand φ in a power series in ik (i is introduced for convenience),

$$\varphi = \sum_{n=0}^{\infty} (ik)^n \frac{\varphi_n}{n!}. \tag{15}$$

We see that in place of equation (12) we obtain

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\varphi_0 = 0, \tag{16a}$$

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\varphi_1 = 0, \tag{16b}$$

and, in general,

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\varphi_n = n(n-1)\varphi_{n-2}. \tag{16c}$$

The boundary condition (13) is satisfied separately by each φ_n. Equation (14) becomes

$$\varphi_0 \underset{\substack{x\to\infty \\ y\to\infty}}{\longrightarrow} 1, \tag{17a}$$

$$\varphi_1 \underset{\substack{x\to\infty \\ y\to\infty}}{\longrightarrow} (x\cos\alpha + y\sin\alpha), \tag{17b}$$

etc. An adequate justification of results, which we have only made plausible, is given in *Morse* and *Feshbach* [1954]. We will need only the first two terms in the series of φ_n.

We make the observation that $\varphi_0 = 1$ will satisfy the boundary conditions and the differential equation. The boundary conditions (17b) and (13), together with the differential equation (16b), are the same as those for the problem of the steady two-dimensional flow of a liquid past a group of objects with a uniform flow at infinity. All the potentials we calculate for various problems will be φ_1.

IV. Typical problems

It will be convenient to do a number of problems involving groups of circular cylinders. For this purpose we use the techniques of the theory of functions, making use of a complex potential [*Lamb*, 1932]. In all problems except part *B* the velocity at infinity is unity.

A. One cylinder in a uniform flow—As a preliminary we put a unit-strength dipole at infinity and calculate the flow past a circular cylinder of radius R. The complex potential is

$$w = ze^{-i\alpha} + \frac{R^2 e^{i\alpha}}{z}. \tag{18}$$

The complex velocity, dw/dz, is

$$\frac{dw}{dz} = e^{-i\alpha} - \frac{R^2 e^{i\alpha}}{z^2} = V_x - iV_y. \tag{19}$$

(In section IV we put $z = x + iy$.) The potential is seen to be real on the cylinder, $z = Re^{i\mu}$, so the cylinder is a streamline. For z large, the velocity vector points in the direction (cos α, sin α), where α is measured from the x axis counterclockwise and has unit magnitude. Consequently the function $w(z)$ given by equation (18) correctly describes the flow past a circular cylinder of a fluid with unit velocity at infinity.

On the surface of the cylinder the potential is

$$w = 2R\cos(\mu - \alpha), \tag{20}$$

and the velocity, which is tangent to the cylinder, is

$$V = \frac{dw}{Rd\mu} = -2\sin(\mu - \alpha). \tag{21}$$

We have chosen positive V to be in the direction of increasing μ along the cylinder.

B. The image of a dipole in a cylinder—Let us consider the flow produced by a dipole of orientation α located at a point $\rho e^{i\beta}$. It has an image of orientation $(\alpha + 2\beta + \pi)$ and relative strength $(R/\rho)^2$ located at $(R^2/\rho)e^{i\beta}$. The complex potential is

$$w = \frac{e^{i\alpha}}{z - \rho e^{i\beta}} - \frac{R^2}{\rho^2}\cdot\frac{e^{i(\alpha+2\beta)}}{\left(z - \frac{R^2}{\rho}e^{i\beta}\right)}. \tag{22}$$

If we put $z = Re^{i\mu}$, we find that the imaginary part of w is independent of μ, so that the surface

of the cylinder is a streamline. Thus

$$w(Re^{i\mu}) = \frac{R\cos(\alpha+\mu+2\beta) - \rho\cos(\alpha+\beta)}{\rho^2 - 2\rho R\cos(\mu-\beta) + R^2} + \frac{e^{i(\alpha+\beta)}}{\rho}, \quad (23)$$

and the velocity, which is tangent to the cylinder is found to be

$$V = [-R\rho^2\{\sin(\alpha+\mu+2\beta) + \sin(\mu-\alpha-2\beta) + \sin(\mu+\alpha)\} + 2\rho R^2\sin(\alpha+3\beta) - R^3\sin(\alpha+\mu+2\beta)]/[R^2 - 2\rho R\cos(\mu-\beta) + \rho^2]^2. \quad (24a)$$

Assuming that R/ρ is small, we find

$$V \cong -\frac{1}{R}\sin(\alpha+\mu+2\beta). \quad (24b)$$

In this section the dipole strength is equal to unity, instead of the velocity at infinity. This fact accounts for the apparent difference in units between equation (20) and equation (24). The results of sections *A* and *B* will now enable us to calculate flows past more complicated geometrical arrays of cylinders.

C. Two cylinders—We begin by consideration of the flow past two circles at $z = 0$ and $z = d$ with radii R_1 and R_2 respectively, a problem for which it is possible also to produce an exact solution. By the method of images we obtain for the potential

$$w = ze^{-i\alpha} + \frac{R_1^2 e^{i\alpha}}{z} + \frac{R_2^2 e^{i\alpha}}{z-d} - \frac{R_1^2}{d^2}\frac{R_2^2 e^{-i\alpha}}{\left(z - \frac{R^2}{d}\right)} - \frac{R_2^2}{d^2}\frac{R_1^2 e^{i\alpha}}{\left(z - d + \frac{R^2}{d}\right)} + \cdots. \quad (25)$$

The dots indicate the omission of images of images, etc. We calculate the velocity along the circle at $z = 0$, assuming that $(R_1/d)^2$ and $(R_2/d)^2$ are small in comparison with one, to be

$$V = -2\left[\sin(\mu-\alpha) - \left(\frac{R_2}{d}\right)^2\sin(\alpha+\mu)\right], \quad (26)$$

for a flow of unit velocity at infinity.

An exact solution to this problem can be found by the use of functions related to elliptic functions. (This problem is solved in *Morse* and *Feshbach* [1954], but no mention is made of this.) We begin by consideration of the mapping function

$$z = ia\tan\frac{\zeta}{2}, \quad (27)$$

which maps the region exterior to two cylinders onto the interior of a rectangle. Corresponding points are indicated in Figure 1. We note in particular that the point $z = \infty$ is mapped on $\zeta = \pi$ and the circles onto the portions of the lines $\zeta = i\xi_2$ and $\zeta = -i\xi_1$, between $\zeta = 0$ and 2π.

In the z plane we have the condition that on the two circles Im w = constant, that is, the circles are streamlines and $w \to ze^{-i\alpha}$ as $z \to \infty$, a uniform flow with an inclination α at infinity. In the ζ plane the point $z = \infty$ corresponds to $\zeta = \pi$. Near this point the potential has the form

$$w \underset{\zeta\to\pi}{\to} -\frac{2aie^{-i\alpha}}{\zeta - \pi}, \quad (28)$$

or from equation (26)

$$w \underset{z\to\infty}{\to} ze^{-i\alpha}, \quad (29)$$

for a flow of unit velocity. The lines $\zeta = i\xi_2$ and $-i\xi_1$ are streamlines and the potential must have a real period of 2π, since the line $\zeta = i\xi$ and $\zeta = i\xi + 2n\pi$ map onto the same line in the z plane.

We can now make a description of the potential. The principle of reflective symmetry of sources in streamlines tells us that the potential must have a double array of poles, half of residue $-2aie^{-i\alpha}$ at points congruent to $(\pi, 0)$ mod $[2n\pi, im(\xi_2, \xi_1)]$, and the other half of residue $+2aie^{-i\alpha}$ at points congruent to $(\pi, -2i\xi_1)$ mod $[2\pi, im(\xi_2+\xi_1)]$.

From this description we conclude that the solution is just the sum of two Z functions [*Whittaker* and *Watson*, 1940]. The Z function

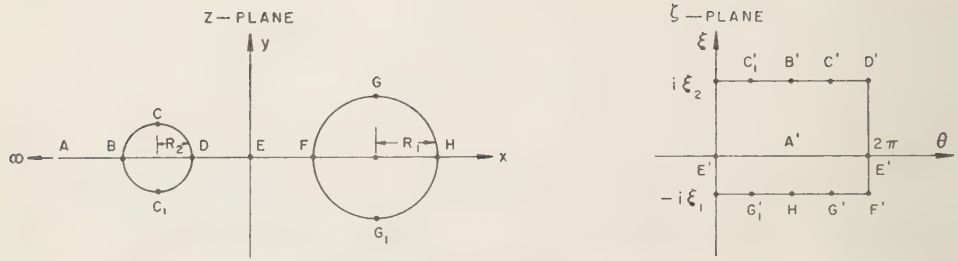


FIG. 1—The mapping of the part of a plane exterior to two circles onto the interior of a rectangle

is closely related to the elliptic functions, having one period and poles on a restangular lattice all with the same residue, 1. If the period of the function is 2π, the poles are at all points congruent to $(0, iK')$ mod $(2\pi, 2miK')$ where K' is a parameter of the Z function which is at our disposal. Putting $K' = \xi_2 + \xi_1$, we see that the solution to our problem is

$$w = -2ia\{e^{-i\alpha}Z[\zeta + \pi - i(\xi_2 + \xi_1)] - e^{i\alpha}Z[\zeta + \pi - i(\xi_2 - \xi_1)]\}, \tag{30}$$

for a source of unit velocity.

The boundary condition on the two cylinders is equivalent to

$$\left.\frac{dw}{d\zeta}\right|_{\xi=\xi_2,\,\xi_1} = \text{real}. \tag{31}$$

This condition is easily verified. We have

$$\left.\frac{dw}{d\zeta}\right|_{\xi=\xi_2} = -2ia[e^{-i\alpha}dn^2(\theta + \pi - i\xi_1) - e^{i\alpha}dn^2(\theta + \pi + i\xi_1)], \tag{32}$$

or

$$\left.\frac{dw}{d\zeta}\right|_{\xi=\xi_2}$$

is real. Moreover, we have

$$\left.\frac{dw}{d\zeta}\right|_{\xi=-\xi_1} = -2ia\{e^{-i\alpha}dn^2[\theta + \pi - i(\xi_2 + 2\xi_1)] - e^{+i\alpha}dn^2(\theta + \pi - i\xi_2)\}. \tag{33}$$

Thus we obtain, since $dn(u) = -dn(u - 2iK')$,

$$\left.\frac{dw}{d\zeta}\right|_{\xi=-\xi_1} = -2ia[e^{-i\alpha}dn^2(\theta + \pi + i\xi_2) - e^{i\alpha}dn^2(\theta + \pi - i\xi_2)]. \tag{34}$$

We see that on $\xi = -\xi_1$, $dw/d\zeta$ is real.

We examine the behavior of w in the vicinity of $\zeta = \pi$. Z has a pole with residue 1 there, so that we find

$$w \underset{\zeta\to\pi}{\longrightarrow} -\frac{2iae^{-i\alpha}}{\zeta - \pi}, \tag{35}$$

or

$$w \underset{z\to\infty}{\longrightarrow} ze^{-i\alpha}. \tag{36}$$

Thus at infinity there is a uniform flow with inclination α. The potential given by equation (30) is the same as equation (25) and either form may be the more convenient, depending on the possible application. It can be shown, by means of some tedious algebra, that to the same order of approximation, equations (26) and (34) agree.

D. Three cylinders in a line—We present approximate solutions to more complicated problems, using the method of images. We consider first three cylinders of radius R located at $z = -d, 0, d$. The potential is, from equation (25),

$$w = ze^{-i\alpha} + R^2e^{i\alpha}\left(\frac{1}{z-d} + \frac{1}{z} + \frac{1}{z+d}\right) - \frac{R^4e^{-i\alpha}}{d^2}\left\{\frac{1}{z - d + \frac{R^2}{d}} + \frac{1}{z - d + \frac{R^2}{2d}} + \frac{1}{z - \frac{R^2}{d}} + \frac{1}{z + \frac{R^2}{d}} + \frac{1}{z + d - \frac{R^2}{d}} + \frac{1}{z + d - \frac{R^2}{2d}}\right\} + \cdots. \tag{37}$$

The three dots indicate the omission of images of images. We find the velocity at the surface

of the center piling to first order in $(R/d)^2$ as a function of μ, the angular position with respect to the x axis, to be

$$V = -2\left[\sin(\mu - \alpha) - 2\left(\frac{R}{d}\right)^2 \sin(\alpha + \mu)\right] + \cdots . \quad (38)$$

The velocity on one of the side cylinders is approximately

$$V = -2\left[\sin(\mu - \alpha) - \frac{5}{4}\left(\frac{R}{d}\right)^2 \sin(\alpha + \mu)\right]. \quad (39)$$

E. Four cylinders on a square—A second problem is that of four smaller circles of radius r located on the four corners of a square and a larger circle of radius R located at the center of the square. If the diagonal of the square is $2d$, and the center circle is at the origin, the potential is, from equation (24),

$$\begin{aligned} w = {} & ze^{-i\alpha} + e^{i\alpha}\left[\frac{r^2}{z-d} + \frac{r^2}{z-id} + \frac{r^2}{z+d} + \frac{r^2}{z+id} + \frac{R^2}{z}\right] \\ & - e^{-i\alpha}\left[\frac{R^2r^2}{d^2}\left\{\frac{1}{z - \frac{R^2}{d}} - \frac{1}{z - i\frac{R^2}{d}} + \frac{1}{z + \frac{R^2}{d}} - \frac{1}{z + i\frac{R^2}{d}} + \frac{1}{z - d + \frac{r^2}{d}} \right.\right. \\ & \left. - \frac{1}{z - i\left(d - \frac{r^2}{d}\right)} + \frac{1}{z + \left(d - \frac{r^2}{d}\right)} - \frac{1}{z + i\left(d - \frac{r^2}{d}\right)}\right\} + \frac{r^4}{d^2}\left\{\frac{1}{z - d + \frac{r^2}{2d}} \right. \\ & \left.\left. - \frac{i}{z - d + \frac{r^2\sqrt{i}}{\sqrt{2}d}} + \frac{i}{z - d - \frac{r^2}{d}\sqrt{\frac{i}{2}}} + \cdots\right\}\right] + \cdots . \end{aligned} \quad (40)$$

The omissions can easily be filled in. To the first order in $(r/d)^2$, we find for the velocity on the center circle

$$V = -2\sin(\mu - \alpha), \quad (41)$$

the shielding effects having just cancelled. The velocity around the circle centered at $z = d$ is approximately

$$V = -2\left[\sin(\mu - \alpha) - \left(\frac{r^2}{4d^2} + \frac{R^2}{d^2}\right)\sin(\alpha + \mu)\right]. \quad (42)$$

F. Six cylinders on a rectangular grid—The final problem to be considered is that of six circles on a rectangular grid. The radius of the circles is R, and the centers are at $(0, 0)$, $(d, 0)$, $(2d, 0)$, $(0, id_2)$, (d_1, id_2), and $(2d_1, id_2)$. The potential is

$$\begin{aligned} w = {} & ze^{-i\alpha} + R^2e^{i\alpha}\left[\frac{1}{z} + \frac{1}{z-d_1} + \frac{1}{z-2d_1} + \frac{1}{z-id_2} + \frac{1}{z-d_1-id_2} + \frac{1}{z-2d_1-id_2}\right] \\ & - R^4e^{-i\alpha}\left[\frac{1}{d_1^2\left(z - \frac{R^2}{d_1}\right)} + \cdots\right]. \end{aligned} \quad (43)$$

The velocity on the circle at $(0, 0)$ is, to the same approximation as previously employed,

$$\begin{aligned} V = {} & -2\left[\sin(\mu - \alpha) - \left\{\frac{R^2}{d_1^2} + \frac{R^2}{4d_1^2} - \frac{R^2}{d_2^2} + \frac{R^2(d_1^2 - d_2^2)}{(d_1^2 + d_2^2)^2} \right.\right. \\ & \left. + \frac{R^2(4d_1^2 - d_2^2)}{(d_1^2 + d_2^2)^2}\right\}\sin(\alpha + \mu) - 2R^2d_1d_2\left\{\frac{1}{(d_1^2 + d_2^2)^2} \right. \\ & \left.\left. + \frac{2}{(4d_1^2 + d_2^2)^2}\right\}\cos(\mu + \alpha)\right]. \end{aligned} \quad (44)$$

For the circle at $(d_1, 0)$, we obtain for the velocity approximately

$$V = -2\left[\sin(\mu - \alpha) - \left\{2\frac{R^2}{d_1^2} - \frac{R^2}{d_2^2} + \frac{2(d_1^2 - d_2^2)R^2}{(d_1^2 + d_2^2)^2}\right\}\sin(\alpha + \mu)\right]. \quad (45)$$

The velocity around any of the others is easily found by a suitable choice of α.

V. Calculation of forces

We can show that all the approximate velocity distributions are the same as they would be if the other pilings were absent, except that the stagnation points are moved and the velocity magnitude is increased or diminished, depending upon the details of the geometry. We note that all the velocity distributions can be written in the form

$$V = -2[\sin(\mu - \alpha) - \delta \sin(\alpha + \mu) - \epsilon \cos(\alpha + \mu)], \quad (46)$$

where δ and ϵ are small quantities depending upon the details of the flow. Let us denote the location of the stagnation point by μ_s and solve equation (46) for μ_s,

$$0 = \sin(\mu_s - \alpha) - \delta \sin(\alpha + \mu_s) - \epsilon \cos(\alpha + \mu_s). \quad (47)$$

Now we have approximately that

$$\mu_s = \mu_0 + \delta\mu_\delta + \epsilon\mu_\epsilon. \quad (48)$$

Substituting equation (48) into (47) and retaining only the linear term in δ and ϵ, we find

$$\mu_0 = \alpha, \quad (49)$$

$$\mu_\delta = \sin 2\alpha, \quad (50)$$

$$\mu_\epsilon = \cos 2\alpha. \quad (51)$$

If we put

$$\nu = \mu - \mu_s \quad (52)$$

into equation (46), we find

$$V = -2 \sin \nu(1 - \delta \cos 2\alpha + \epsilon \sin 2\alpha). \quad (53)$$

We see by comparison with equation (21) that the velocity distribution has been changed only as to the over-all magnitude and the location of the stagnation point.

The usual engineering method for the calculation of wave forces on a single vertical circular cylinder is to employ the formulas of *Morison* and others [1954],

$$F = \rho\pi R \int_0^{\text{surface}} C_D u\,|u|\,dz + \rho\pi R^2 \int_0^{\text{surface}} G_I \frac{\partial u}{\partial t}\,dz. \quad (54)$$

The force is the sum of two integrals. The first integral is called the drag force and is a direct generalization of the usual steady-state relationship between force and velocity. The second integral is called the inertial force, which is due to the accelerations of the water. The function u is the undisturbed particle velocity (that is, if the pile were not present). The drag coefficient C_D and the inertial coefficient C_I are experimentally determined functions. While the representation of the force in this fashion is undoubtedly too simplified, recent measurements indicate that it is a useful first approximation.

Now the point of the calculation was to show that whether there was one cylinder present or many, provided only that the geometrical arrangement was correct, the velocity distribution around the cylinder was changed only by a relocation of the stagnation points and a change in the magnitude of the velocity. Thus we could, after our calculation, construct a wave of known height and direction of incidence on the cylinder which would produce exactly the same velocity distribution on it alone as it would have experienced because of its position in the group of cylinders. Thus in calculating the force phenomenologically, by use of drag and inertial coefficients, we would employ slightly different drag and inertial coefficients and different velocities and nonconvective accelerations. It is true, however, that measured values of drag and inertial coefficients show rather slow variation with the magnitude of the velocity, so that to a first approximation it is consistent to correct only the velocities and nonconvective accelerations. So we see that there is only one geometrical factor to be used in correcting the inertial force and that its square is used to modify the drag force. These correction factors are listed for the cases we have considered in Table 1.

VI. Concluding remarks

We review the various approximations used in developing the theory to see, in a general way, how they affect its validity. We have assumed that the presence of a detached boundary layer,

or other local departure from potential flow, is responsible for the force on a single piling. Then we can calculate the interaction between pilings by the methods of classical hydrodynamics, neglecting viscosity. This assumption cannot be strictly correct, in view of the well-known phenomenon of the shedding of vortices in the steady flow past a cylinder. The vortices are propagating effects of the local disturbed boundary layer, which is a direct consequence of the viscosity. The results of the calculations in all cases were that the velocity distribution was unchanged, except for multiplication by a numerical factor and shifting of the location of the stagnation points. Then for engineering calculations it is necessary only to modify the drag and inertial coefficients by geometrical factors which reflect the change in the velocity due to the presence of the other cylinders, since the drag and inertial coefficients themselves are only slowly varying functions of the velocity and local acceleration. It is common practice to consider them to be constants.

Three geometrical assumptions have been made. First, we have assumed that we can use the linearized wave theory, making the approximation that the wave height is a small fraction of the wavelength. This is not so serious, since we treat the one-cylinder problem exactly by using experimental results. Second, we have made a static approximation, assuming that the over-all dimensions of the obstacles are a small fraction of a wavelength. It is clear that the retardation, which is included if higher order terms are taken, would upset the calculation, since it is quite unreasonable to expect such simple results for the velocity distribution. Finally, we have assumed that the cylinder radii are small in comparison to their separation. This assumption is also necessary for the validity of the calculations, as can be seen by looking at the exact solution for the problem of two cylinders or carrying out one more step in the various approximate solutions. Fortunately, these approximations are often satisfied in cases of practical interest.

References

Lamb, Horace, *Hydrodynamics*, Dover Publications, VI Edition, New York, 1932.

Morison, J. R., J. W. Johnson, and M. P. O'Brien, Experimental studies of forces on piles, *Proc. Fourth Conf. Coastal Eng.*, held in Chicago, Illinois, Oct. 1953. Printed by the Eng. Foundation, Univ. of California, Berkeley, California, p. 340, 1954

Morse, Philip M., and Herman Feshbach, *Methods of Theoretical Phys.*, McGraw-Hill, New York, 1954.

Whittaker, E. T., and G. N. Watson, *Modern Analysis*, Cambridge Univ. Press, Cambridge, 1940.

(Manuscript received September 29, 1958; presented at the Thirty-Ninth Annual Meeting, Washington, D. C., May 6, 1958.)

List of symbols

a—distance used in mapping given by equation (27)
A—$-dH/2i\omega$
d—various distances
d_1—a distance
d_2—a distance
D—depth
g—acceleration of gravity
H—wave height
i—$\sqrt{-1}$
k—$2\pi/\lambda$
K'—arbitrary constant used with elliptic functions
(n_x, n_y)—x and y components of normal vector
p—pressure
r—radius of a cylinder
R—radius of a cylinder
R_1—radius of a cylinder
R_2—radius of a cylinder
t—time
u—particle velocity
V—particle velocity on various cylinder surfaces
w—complex potential
x—coordinate
y—coordinate
z—coordinate (except in section IV)
z—$x + iy$ (only in section IV)
Z—Zeta function
α—angle giving direction of waves
β—angular location of a certain point
δ—small quantity
ϵ—small quantity
ζ—$2 \tan^{-1} (z/ia)$
θ—real part of ζ
λ—wavelength
μ—angle locating positions on various circles
μ_s—position of stagnation point

μ_0—α
μ_δ—sin 2α
μ_ϵ—sin 2ϵ
ν—$\mu - \mu_\epsilon$
ξ—imaginary part of ζ
ξ_1—specific value of ξ
ξ_2—specific value of ξ
ρ—density of water (except in section IV)
ρ—distance (only in section IV)
φ—Φ modified to omit z dependence, equation (8)
$\bar{\varphi}$—velocity potential
Φ—velocity potential without time dependence, equation (6)
φ_n—equation (15)
ω—angular frequency of the waves
Ω—gravitation potential

A Determination of the Coefficient J of the Second Harmonic in the Earth's Gravitational Potential from the Orbit of Satellite 1958 β_2

MYRON LECAR*, JOHN SORENSON*, AND ANN ECKELS*

Theory and Analysis Branch, Project VANGUARD
U. S. Naval Research Laboratory
Washington, D. C.

Abstract—From the secular changes in the longitude of the node and the argument of the perigee, the coefficient J in the gravitational potential function

$$U = \frac{\mu}{r}\left[1 + J\left(\frac{a_e}{r}\right)^2\left(\frac{1}{3} - \sin^2\phi\right) + \frac{K}{30}\left(\frac{a_e}{r}\right)^4(3 - 30\sin^2\phi + 35\sin^4\phi)\right]$$

was determined to be $(1.6232 \pm .0005) \times 10^{-3}$. If the bounding equigeopotential surface of the earth is assumed to be an ellipsoid of revolution, this value of J implies a flattening f of $1/298.32 \pm .05$.

I. INTRODUCTION

The determination of the coefficients of the perturbing function, that is, the non-central part of the gravitational potential was based on the orbit of satellite 1958 β_2. Approximately one hundred and eighty Minitrack observations were used, covering a period of ninety days. This period coincides with the period of the argument of perigee and was so chosen to average the effects of long-period perturbations. In addition, the observations were selected from the approximately eight hundred available observations with the criterion of using only nearly overhead passes (within five degrees of zenith), thus minimizing errors due to ionospheric refraction. Since the orbit determination is such an integral part of this study, it may be well to devote some time to a discussion of the 'orbit elements.'

The parameters of a satellite orbit may be defined as follows. The planar elements of the ellipse are the semi-major axis a, and the eccentricity e. In addition, we indicate the position of the satellite in the orbital plane by one of three angles. The true anomaly θ is the angle subtended at the focus by the line of apsides and the radius vector (originating at the focus) to the satellite. The eccentric anomaly E is the angle subtended at the center of the ellipse by the line of apsides, and the radius vector originating at the center of the ellipse, and terminating at the intersection of a line—perpendicular to the line of apsides, and passing through the satellite—with a sphere of radius a concentric with the ellipse. The mean anomaly M is introduced through Kepler's equation $M = E - e \sin E$. The three angles are measured positive from the perigee π in the direction of the satellite motion, as illustrated in Figure 1.

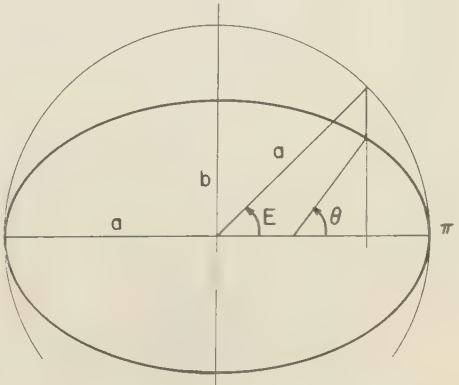


FIG. 1—Ellipse parameters

* Present addresses respectively; Physics Dept., Yale University, New Haven, Conn.; Mathematics Dept., Purdue University, West Lafayette, Indiana; and Theoretical Division, National Aeronautics and Space Administration, Washington 25, D. C.

We may alternatively replace a by the anomalistic (perigee-to-perigee) period P or the mean motion n which is equal (in radians) to $2\pi/P$.

Further, we fix the plane and orientation of the ellipse with three additional parameters. The inclination i is the angle between the plane of the orbit and the earth's equatorial plane. The argument of perigee ω is measured from the line of nodes (the line forming the intersection of the orbital and earth's equatorial planes) to the line of apsides. ω is measured positive from the ascending node (the point where the satellite crosses the equator from south to north) to perigee in the direction of satellite motion. Finally, the longitude of the ascending node Ω is measured positive to the east in the plane of the equator from the vernal equinox to the ascending node.

A reference frame, with the X-Y plane as the plane of the equator, the X-axis going through the vernal equinox, the Z-axis coincident with the earth's axis of rotation and directed north, and the Y-axis forming a right-handed system, is considered to be an inertial frame. The vernal equinox is taken at 1958.0 and nutation of the earth's axis is neglected.

The effect of the non-central components of the earth's gravitational potential produces periodic and secular variations of the orbit elements. Their instantaneous values are called 'osculating elements.' We may also define 'mean elements' as those values of a, e, and i, that one would assign to a rigid ellipse (a, e, and i, constant) precessing with the secular rates in Ω and ω, to effect a best least-squares fit of the position of a satellite—travelling in the rigid ellipse with average n—and the actual position of the satellite.

Recently, a number of authors [*Blitzer*, 1957; *Blitzer* and *Wheelon*, 1957; *Blitzer* and others, 1957; *Garfinkel*, 1957; *Herget*, 1957; *Krause*, 1956; *Okhocimski* and others, 1957] have investigated the perturbations of a satellite orbit resulting from the earth's oblateness. In particular, we have made extensive use of Herget and Musen's General Theory of Oblateness Perturbations and the work of Krause.

It is found that, neglecting air resistance (which for this satellite, 1958 β_2, amounts to a shortening of the semi-major axis of less than one part in 10^6 per day), only Ω and ω have secular changes. For satellite 1958 β_2, the changes in Ω and ω amount to about three and four degrees per day, respectively. The line of nodes regresses (moves in a direction opposite to that of the satellite motion) and the line of apsides moves with the satellite, as is illustrated schematically in Figure 2.

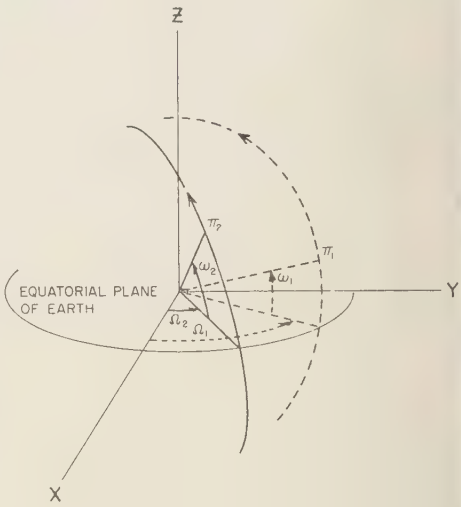


FIG. 2—Rotation of the line of nodes and the line of apsides

One would expect over a long period, $\dot{\Omega}$ and $\dot{\omega}$ could be measured with sufficient precision to be useful in inferring the non-central components of the earth's gravitational field. This, in fact, was suggested by *Krause* [1956], and others [*Blitzer* and others, 1957; *Okhocimski* and others, 1957], and in June of this year, a preliminary determination of the flattening was announced by *O'Keefe* [1958] of the U. S. Army Map Service. His determination of f from 1958 β_2 was 1/298.3, which is seen to agree very well with the value determined here.

II. Geodetic relationships

Taking the bounding equi-potential surface of the earth to be an ellipsoid of revolution, the following relations can be derived (to order f^2). Let the geopotential function ψ be given by

$$\psi = U + \tfrac{1}{2}\omega_e^2 r^2 \cos^2\phi \qquad (1)$$

The gravitational potential U can be expressed in spherical harmonics as

$$U = \frac{\mu}{r}\left[1 + \left(\frac{a_e}{r}\right)^2 J\left(\frac{1}{3} - \sin^2\phi\right) + \left(\frac{a_e}{r}\right)^4 \frac{K}{30}(3 - 30\sin^2\phi + 35\sin^4\phi) + \cdots\right] \quad (2)$$

where ω_e is the angular velocity of the earth about its polar axis,

r is the radius vector,

ϕ is the geocentric latitude,

a_e is the earth's equatorial radius.

With our initial assumption, equations (1) and (2) yield

$$J = f - \frac{1}{2}\rho + \frac{9}{14}\rho f - \frac{1}{2}f^2 + 0(f^3) \quad (3a)$$

$$K = 3f^2 - \frac{15}{7}\rho f + 0(f^3) \quad (3b)$$

where $\rho = \omega_e^2 a_e^3/\mu$ and f is the flattening or ellipticity.

We also find that

$$g_e = \frac{\mu}{a_e^2}\left[1 + J + \frac{K}{2} - \rho\right] \quad (4a)$$

and

$$g = g_e[1 + \beta\sin^2\phi + \gamma\sin^2 2\phi] \quad (4b)$$

where g_e is the value of g on the equator,

ϕ = the geodetic latitude,

$$\beta = -f + \frac{5}{2}\rho + \frac{15}{4}\rho^2 - \frac{26}{7}\rho f,$$

$$\gamma = \frac{1}{8}f^2 - \frac{5}{8}\rho f.$$

Further, substituting equations (3a) and (3b) in (4a), we obtain

$$\mu = \frac{\omega_e^2 a_e^3(\frac{3}{7}f + \frac{3}{2}) + g_e a_e^2}{(1 + f + f^2)} \quad (5)$$

In this paper, the following quantities were taken to be exact:

a_e = 6378270 meters

$\omega_e = .7292115 \times 10^{-4}$ rad/sec

g_e = 978.038 cm/sec^2 for $f = 1/297$

To determine μ, equation (4a) was altered* to read

$$g_e = \frac{\mu}{a_e^2}\left[1 + J + \frac{K}{2} - \rho - A\right] \quad (6)$$

where A is the mass of the atmosphere divided by the total mass of the earth (atmosphere included). A was taken to be $.68 \times 10^{-6}$ [*Herrick* and others, 1957]. Also the acceleration at the equator was taken to be

$$g_e = 978.038\left[1 + \frac{1}{3}\left(f - \frac{1}{297}\right)\right] \quad (7)$$

and was determined by repeated iteration. The correction for f is due to the fact that the preponderance of surface measurements of g is in the middle latitudes ($\sin^2\phi = \frac{1}{3}$).

These relationships enable us to determine μ once f is known. They also provide a relationship between J and K, which for values of f ranging from 1/297 to 1/299, is seen to be linear. In particular, we use

$$K = (-1.1818 \times 10^{-5}) + (1.2730 \times 10^{-2})J \quad (8)$$

Graphs of J vs. f and J vs. K are shown in Figures 3 and 4. In addition, it is often convenient in the computation of orbits to use canonical units of length and time such that μ becomes numerically equal to unity. A table of μ and T (canonical unit of time) vs. f, where the unit of length is taken as a_e, is also presented (Table 1).

$$T = \sqrt{\frac{A_e^3}{\mu}}, \qquad A_e = 637827000 \text{ cm}$$

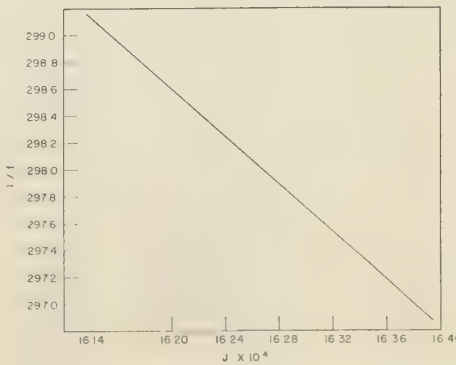


FIG. 3—Oblateness (f) vs. J

* Note that to the extent that the atmosphere can be considered as being composed of confocal homogeneous shells, its mass will have no effect on earth-bound measurements. This, of course, does not hold for the satellite.

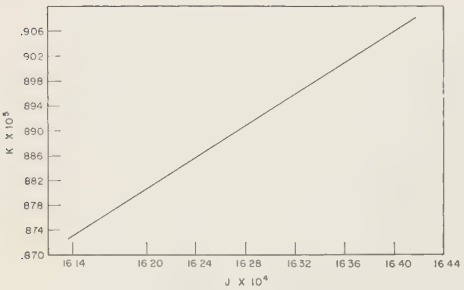


FIG. 4—J vs. K

III. Orbit perturbations to order J

If terms containing K in the potential are disregarded, Krause has found that the following relationships hold for the secular motion of the longitude of the node and the argument of the perigee:

$$\frac{\dot{\Omega}}{2\pi} = \frac{-\cos i}{p^2} J \tag{9}$$

$$\frac{\dot{\omega}}{2\pi} = \frac{2(1 - \frac{5}{4}\sin^2 i)}{p^2} J \tag{10}$$

where $p = a(1 - e^2)$ and $\dot{\Omega}$ and $\dot{\omega}$ are expressed in radians/period. Differentiating equations (8) and (9), one finds the absolute magnitude of the relative errors of J to be

$$\left|\frac{dJ}{J}\right| = \left[\left(\frac{d\dot{\Omega}}{\dot{\Omega}}\right)^2 + \tan^2 i(di)^2 + \frac{49}{4}\left(\frac{da}{a}\right)^2 + \frac{1}{4}\left(\frac{d\mu}{\mu}\right)^2 + \frac{16e^4}{(1-e^2)^2}\left(\frac{de}{e}\right)^2\right]^{\frac{1}{2}} \tag{11}$$

$$\left|\frac{dJ}{J}\right| = \left[\left(\frac{d\dot{\omega}}{\dot{\omega}}\right)^2 + \frac{25}{4}\frac{(\sin i \cos i)^2}{(1 - \frac{5}{4}\sin^2 i)^2}(di)^2 + \frac{49}{4}\left(\frac{da}{a}\right)^2 + \frac{1}{4}\left(\frac{d\mu}{\mu}\right)^2 + \frac{16e^4}{(1-e^2)^2}\left(\frac{de}{e}\right)^2\right]^{\frac{1}{2}} \tag{12}$$

The terms in μ enter in the following way. $\dot{\Omega}$ and $\dot{\omega}$ were originally determined in radians/day. To first order, the period is given by $2\pi a^{3/2}/\sqrt{\mu}$, and the relative error in the period is

$$\frac{dP}{P} = \frac{3}{2}\left(\frac{da}{a}\right) - \frac{1}{2}\left(\frac{d\mu}{\mu}\right)$$

To evaluate equations (9) to (12), the following orbit elements were used:

$a = 1.361833 \pm .000008$ earth's equatorial radii
$e = .1904 \pm .0003$
$i = .5977 \pm .0002$ radians

These values were derived from fourteen determinations of the orbit elements, each covering an arc of approximately six days. Assuming a to decrease linearly with time and i and e to remain constant, a least-squares fit was made of the elements presented in Table 2.

$\dot{\Omega}$ and $\dot{\omega}$ were obtained by differencing the first and last values as follows:

Epoch days	Ω (*radians*)	ω (*radians*)
9.51875	$2.2137 + 2\pi \pm .00035$	$2.9307 \pm .00016$
94.51875	$4.0276 \pm .00023$	$3.3911 + 2\pi \pm .00013$

$\dot{\omega} = .076808 \pm .000008$ radians/day = $.0071621 \pm .0000008$ radians/period*.

$\dot{\Omega} = -.052581 \pm .000004$ radians/day = $-.0049030 \pm .0000004$ radians/period.

Then from equations (9) and (11) we find

$$J = (1.6260 \pm .0006) \times 10^{-3}$$

TABLE 1—μ *and* T *vs.* f

$\frac{1}{f}$	$\mu(10^{20}$ cm^3/sec$^2)$	T(sec)
297.0	3.9861402	806.8226
297.1	3.9861432	806.8223
297.2	3.9861462	806.8220
297.3	3.9861492	806.8217
297.4	3.9861523	806.8214
297.5	3.9861553	806.8211
297.6	3.9861583	806.8208
297.7	3.9861613	806.8205
297.8	3.9861643	806.8202
297.9	3.9861673	806.8199
298.0	3.9861703	806.8196
298.1	3.9861733	806.8193
298.2	3.9861764	806.8190
298.3	3.9861794	806.8187
298.4	3.9861824	806.8184
298.5	3.9861854	806.8180
298.6	3.9861884	806.8177
298.7	3.9861913	806.8174
298.8	3.9861943	806.8171
298.9	3.9861973	806.8168

* The period is determined from the relationship $P = 2\pi a^{3/2}$ where P and a are in canonical units.

TABLE 2—*Orbit elements of satellite 1958 β_2*

Epoch (days after launch)*	a (in units of 6378388m)	e	i (radians)	ω (radians)	Ω (radians)	M (radians)
9.51875	1.3619219	.19068	.59755	2.9306910	2.2137827	3.2902109
15.51875	1.3619126	.19044	.59783	3.3911148	1.8985333	5.4235408
27.51875	1.3618835	.19006	.59805	4.3140994	1.2675772	3.4261125
39.51875	1.3618642	.18988	.59815	5.2379634	0.6369483	1.4471124
45.51875	1.3618464	.19018	.59778	5.6999231	0.3209439	3.6079968
57.51875	1.3618230	.19036	.59783	0.3386158	5.9743206	1.6672045
60.51875	1.3618088	.19069	.59777	0.5684311	5.8157352	2.7565842
64.77431	1.3618031	.19068	.59773	0.8965832	5.5909566	0.4873017
70.51875	1.3617915	.19080	.59772	1.3339909	5.2898012	4.3045853
82.51875	1.3617644	.19042	.59754	2.2547897	4.6585044	2.4125217
85.51875	1.3617582	.19048	.59762	2.4842030	4.5014992	3.5139459
88.51875	1.3617541	.19049	.59753	2.7153902	4.3428677	4.6164890
91.51875	1.3617491	.19028	.59788	2.9455980	4.1855980	5.7194140
94.51875	1.3617448	.19022	.59757	3.1762341	4.0276082	0.5408737

* Launch date = 1958 March 17.51875

and from equations (10) and (12)

$$J = (1.6251 \pm .0009) \times 10^{-3}$$

We note, for comparison, that the value of J associated with $f = 1/297$ is 1.6383×10^{-3} which is well outside of the standard deviations of the values found above.

IV. Orbit perturbations to order K

The theory of Krause—Krause's extension of equations (9) and (10) to order K are given below (in Krause's notation, $J = 3K_2$, $K = 10K_4$).

$$-\frac{\dot{\Omega}}{2\pi} = \frac{\cos i}{p^2} J + \frac{\cos i}{p^4}\left(1 + \frac{3}{2}e^2\right)\left(1 - \frac{7}{4}\sin^2 i\right)K \quad (13)$$

$$\frac{\dot{\omega} + \dot{\Omega}\cos i}{2\pi} = \frac{(1 - \frac{3}{2}\sin^2 i)}{p^2} J + \frac{(1 + \frac{3}{4}e^2)(1 - 5\sin^2 i + \frac{35}{8}\sin^4 i)}{p^4} K \quad (14)$$

Using the orbit elements presented in the previous section, and the linear relationship between J and K in equation (8), we find from equations (13) and (14), respectively:

$$J = 1.6237 \times 10^{-3}, \qquad K = 0.885 \times 10^{-5}$$

and

$$J = 1.6251 \times 10^{-3}, \qquad K = 0.887 \times 10^{-5}$$

Redetermination of the orbit and the coefficient J—The methods used to generate orbits at the VANGUARD Computing Center are the method of General Oblateness Perturbations (GOP) and Numerical Integration of the differential equations of motion (NI). Either of the above is used in conjunction with a differential correction program, to effect a best least-squares fit of the observed vs. computed positions of the satellite (actually, angular measures rather than positions are used). All of the preceding was formulated by Dr. Herget, and programmed by the IBM staff of the Center. A comparison of GOP, NI, and the equations of Krause will be presented in the next section.

At this point in the study, it was felt that J was sufficiently determined to warrant a redetermination of the orbit with this new gravitational potential. Approximately two hundred Minitrack observations were selected, with the criterion that the satellite should pass within five degrees of the observer's zenith; this, of course, to minimize ionospheric refraction errors. These observations covered a period of ninety days.

The first such determination was actually made earlier in the study. Using the then current values of J and K (corresponding to $f = 1/297$), the average residuals (difference between observed and computed angular measures) was

TABLE 3—*Determination of the orbit as a function of J*

Run	$J \times 10^3$	Mean residuals	Values after differential correction			Secular rates used by GOP	
			a	e	i	$-\dot{\Omega}$	$\dot{\omega}$
1	1.63810	12.59	1.36182980(10)	0.19040(10)	0.6002(3)	0.053058	0.077503
2	1.62570	5.25	1.36182678(4)	0.19047(6)	0.5983(1)	0.052658	0.076902
3	1.62390	4.92	1.36182638(4)	0.19049(6)	0.5980(1)	0.052599	0.076817
4	1.62324	4.88	0.36182626(4)	0.19049(5)	0.5978(1)	0.052578	0.076785
5	1.62210	4.97	0.36182599(4)	0.19051(6)	0.5978(1)	0.052541	0.076731

In the above, the 'mean residuals' are in milliradians. a is in units of earth's equatorial radius, i in radians, and $\dot{\Omega}$ and $\dot{\omega}$ in radians/day. The numbers in parentheses refer to the standard deviation in the least significant digit.

about thirteen milliradians. Since, for a short arc determination, for example, six days, this quantity usually runs about one milliradian, we had hopes of considerable improvement to the orbit.

In all, five such determinations were made. After each determination a comparison was made between the observed and computed $\dot{\Omega}$, and equation (13) was used, in effect, to differentially correct the oblateness coefficients. The results of these determinations are presented in Table 3 (also see Fig. 5).

The value of J corresponding to minimum residuals is

$$J = 1.6232 \pm .0005 \times 10^{-3} \qquad (15)$$

where the standard deviation was computed as a function of the standard deviations of the elements of Run No. 4, with the aid of equation (13).

Using equations (8) and (3a), respectively, we find:

$$K = .8845 \pm .0006 \times 10^{-5} \qquad (16)$$

$$f = 1/298.32 \pm .05 \qquad (17)$$

We remind the reader that, to some extent, the value of J is a function of the model chosen for the equi-geopotential surface (Section I). K and f are, of course, strongly dependent on this model.

V. Comparisons of perturbation theories and conclusions

In comparing the various methods of orbit determination used in this study, we encounter the difficult problem of defining the orbit

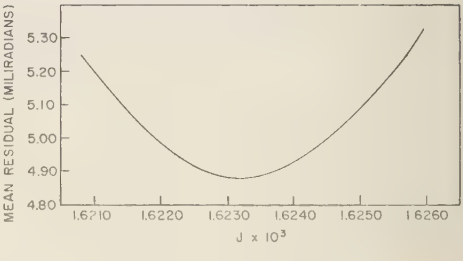


FIG. 5—Average residual vs. J

elements. At the VANGUARD Computing Center, the most frequently used programs are a GOP program and a differential correction program. The elements presented in the previous sections are GOP elements, and we will now illustrate numerically a comparison of GOP elements with osculating and mean elements. The mathematical formulation of GOP will appear in a forthcoming publication by Herget.

The output of the GOP program consists of a set of base elements and the perturbations to these base elements, the latter in the form of Fourier expansions. For example, the planar equations evaluated at apogee and perigee are as follows:

$$\begin{aligned} r(\text{apogee}) &= a(1 + e)(1 + u) \\ r(\text{perigee}) &= a(1 - e)(1 + u) \end{aligned} \qquad (18)$$

where u is in the form of a Fourier Series with arguments E and ω, and, in addition, contains a constant term. For the satellite 1958 β_2, the constant term is usually of the order 10^{-4}. The output of GOP can be evaluated to give instan-

TABLE 4—*GOP elements vs. NI elements*

a		e		i		Time (min)
GOP	NI	GOP	NI	GOP	NI	
1.362437	1.362436	0.190711	0.190711	0.5977955	0.5977955	0
1.362623	1.363316	0.191230	0.191373	0.59813	0.59808	12
1.362482	1.362476	0.190705	0.190703	0.59782	0.59782	24
1.361877	1.361870	0.190372	0.190369	0.59759	0.59759	36
1.361973	1.361966	0.190306	0.190302	0.59766	0.59766	48
1.362163	1.362155	0.190090	0.190087	0.59783	0.59782	60
1.362268	1.362264	0.189896	0.189890	0.59794	0.59794	72
1.362285	1.362285	0.189869	0.189864	0.59796	0.59796	84
1.362226	1.362216	0.190035	0.190033	0.59787	0.59787	96
1.362052	1.362042	0.190281	0.190278	0.59771	0.59771	108
1.361869	1.361862	0.190387	0.190383	0.59758	0.59758	120
1.362206	1.362262	0.190353	0.190363	0.59781	0.59780	Average values

taneous position, and with a seventh-order differencing we have extracted instantaneous velocity. However, for an evaluation over the entire ninety-day arc, this method is obviously cumbersome. Mean elements may be obtained, to a good approximation, by evaluating the u-series at apogee and perigee and solving equations (18).

In the case of Krause's formulas [equations (13) and (14)], we compare the results of using mean elements with using osculating elements at various portions of the orbit and then averaging the resulting $\dot{\Omega}$'s.

Numerical integration, of course, uses osculating elements as starting values. The particular numerical integration program used here has been checked for accumulated errors by integrating a Keplerian orbit for ten days, using an integration interval of one minute. The difference in the osculating elements at the beginning and end of the run is never in excess of one part in 10^6.

A tabulation of osculating a, e, and i equally spaced eleven times around one revolution of the satellite is presented in Table 4. This tabulation was computed by GOP and Numerical Integration. The first values listed for NI were the starting values for the integration. The GOP values are derived from Run No. 4 of the previous section. The first time listed is 1958, May 8.45069 Z.

Equations (13) and (14) were evaluated using the GOP osculating elements, and then the individual determinations of $\dot{\Omega}$ and $\dot{\omega}$ were averaged. This is to be compared with first averaging the osculating elements, and using the average values of the elements in equations (13) and (14). The former method is labelled A in the tabulation below.

	$\dot{\omega}(rad/day)$	$\dot{\Omega}(rad/day)$
A	0.0767580	−0.0527190
B	0.0767581	−0.0527189

Although equations (13) and (14) give the secular rates as a function of osculating elements, we see that to good approximation, mean elements may be used. We attribute the discrepancies of the above values as compared with the observed values to the fact that the above elements cover too small a sample of the ninety day arc.

For a final comparison, we computed mean GOP elements by the method described previously. They are:

$$a = 1.36167233(4)$$
$$e = 0.19040(5)$$
$$i = 0.5978(1).$$

The basic elements from Run No. 4 were:

$$a = 1.36182626(4)$$
$$e = 0.19049(5)$$
$$i = 0.5978(1)$$

The secular rates of $\dot{\Omega}$ and $\dot{\omega}$ as computed by GOP in equations (13) and (14) for mean GOP elements and the observed values are presented below.

	$\dot{\Omega}(rad/day)$	$\dot{\omega}(rad/day)$
GOP	−0.052578	0.076785
Krause	−0.052576	0.076762
Observed	−0.052581	0.076808

The orbit of 1958 β_2 has been well determined due in part to its inherent stability, and in a large measure, to the efficaciousness of Herget's programs and the large number of available Minitrack observations. This enabled the determination of the gravitational potential to the accuracy presented. The resulting value for J, however, is, to some extent, a function of the model assumed for the equi-potential surface, since the relationship derived from this model was used to eliminate K as an independent variable. A number of studies [*Herrick* and others, 1957; *Jeffreys*, 1952; *Jones*, 1952] were available which indicated results in disagreement with this model, and it is anticipated that terms of order K will have to be included to account for the non-spheroidicity of the equi-geopotential surface. However, since satellite determinations, because of the large area sampled, add a new dimension to the study of the potential of the earth, we felt that at the outset it would be advantageous to retain the simpler model. The departures should be small enough to be treated as perturbations, and retention of a simple model will ease the problem of correlating future results.

Acknowledgments—We, first of all, must acknowledge our deep indebtedness to Paul Herget, Director of the Cincinnati Observatory, who, in addition to formulating GOP, NI, the differential correction program, and a host of auxiliary programs, provided encouragement and invaluable personal advice throughout the study. Further, the staff of the VANGUARD Computing Center performed excellent service in programming for the IBM-704 the aforementioned formulations. In particular, we would like to thank Israel Krongold, Bruce Oldfield, Robin Mowlem, Chester Raker, and George Collins for their efforts during this study. Hugh Chrisman of the Laboratory was most cooperative in providing machine time. John O'Keefe and S. Henricksen of the Army Map Service provided many useful references. H. Hauptman of the Laboratory, and J. Frank of American University were of considerable help in interpreting Krause's paper.

We should make mention of the excellent work of J. T. Mengel and his staff of the Minitrack Branch of Project VANGUARD who provided the observations. Finally, we would like to thank J. W. Siry, Head of the Theory and Analysis Branch of Project VANGUARD, who made this study possible.

References

Blitzer, L., Apsidal motion of an IGY satellite *orbit*, *J. Appl. Phys.*, **28**, p. 1362, 1957.

Blitzer, L., and A. D. Wheelon, Oblateness perturbation of elliptical satellite orbits, *J. Appl. Phys.*, **28**, p. 279, 1957.

Blitzer, L., M. Weisfeld, and A. D. Wheelon, Perturbations of a satellite's orbit due to the earth's oblateness, *J. Appl. Phys.*, **27**, 1141–1149, 1957.

Chovitz, B., and I. Fischer, A new determination of the figure of the earth from arcs, *Trans. Am. Geophys. Union*, **37**, 534–545, 1956.

Garfinkel, B., *On the motion of a satellite of an oblate planet*, BRL Report No. 1018, July 1957.

Herrick, S., R. M. L. Baker, Jr., and C. G. Hilton, *Units and constants for geocentric orbits*, Consultants for IGY Satellite Program through the Smithsonian Astrophysical Laboratory, 1957.

Herget, P., *General theory of oblateness perturbations*, American Mathematical Society Symposium, New York University, April, 1957.

Jeffreys, H., *The earth*, Cambridge University Press, 1952.

Jones, H. S., *The Earth as a Planet*, Chapter I, edited by Kuiper, G. P., University of Chicago Press, 1952.

Krause, H. G. L., *"Die Sakularen und Periodischen Storungen der Bahn eines Kuntstlichen Erdsatelliten"*, VII Internationalen Astronautischen Kongress, September, 1956.

O'Keefe, J. A., *"Oblateness of the earth by artificial satellites"*, Harvard College Observatory Announcement Card 1408, June 24, 1958.

Okhocimski, D. Ye., T. M. Eneyev, G. P. Taratynova, and I. M. Yacunski, Orbital perturbations of artificial satellites, *Uspekhi Fiz. Nauk*, 63, Translated by Hope, E. R., Defence Research Board T270R, Canada, 1957.

(Manuscript received October 3, 1958.)

Gravity Measurements between Hazen and Austin, Nevada: A Study of Basin-Range Structure

George A. Thompson

Geophysics Department
Stanford University
Stanford, California

Abstract—From the pendulum base stations at Mystic and Truckee, California, a line of gravimeter stations, with side loops, was extended eastward through Hazen, Fallon, Eastgate, and Austin, Nevada, crossing the area displaced by faults in 1954. Regionally, the Bouguer anomaly is about −160 mgal from Hazen eastward through Fallon to the Stillwater Range. Farther east the values decrease to −215 mgal near Austin, a change that can be accounted for by isostatic compensation for the increase in average elevation of about 1700 ft. Low-density sedimentary and volcanic debris is abundant enough in the intermontane basins and in parts of the mountains to make the average density of the whole elevated land mass abnormally low with respect to the conventional value, 2.7 g/cc. Consequently, 10 to 15 pct of the compensation for this high region is in the superficial materials themselves. The Basin Ranges show no evidence of being individually compensated.

Locally, each of the basins has a negative Bouguer anomaly relative to the adjacent ranges, reflecting thick sedimentary and volcanic fill. Dixie and Fairview Valleys, in the area displaced by faults in 1954, are characterized by local negative anomalies of about 30 mgal, indicating that these valleys contain several thousand feet of lightweight Cenozic sediments and that their bedrock floors lie below sea level. The 1954 faulting is thus only the latest of many displacements that produced not only the visible topographic relief of 5000 ft but also a buried relief of comparable magnitude.

Romney's [1957] seismic studies of the 1954 faulting and *Whitten's* [1957] geodetic measurements agree with direct observations of fault surfaces to indicate a horizontal extension of about 5 ft normal to the trace of the fault. Independent of a strike-slip component of faulting, the region is expanding in area. If the total structural relief was produced by displacements comparable to that of 1954, the extension across Dixie and Fairview Valleys amounts to about 1½ mi in roughly 15 million years, and if this is taken as a fair sample of the Basin and Range Province, the rate of distension in the Province is about 1 ft/century, a rate well within that of historical fault displacements.

At the time of the 1954 faulting, the Bouguer gravity anomaly either did not change or decreased algebraically by an amount no greater than 1.0 mgal.

Introduction—The region under study lies in the Basin and Range Province in west-central Nevada and includes an area broken by faults in 1954 (Fig. 1).

Gravity measurements were undertaken to aid in understanding the structural history. Fault movements associated with the earthquakes of December 16, 1954, provide a dynamic demonstration of basin and range structures in the making. From July to December a series of earthquakes shook a large area centered east of Fallon, and two main shocks on December 16 were accompanied by surface breaks extending discontinuously for sixty miles in a northerly direction and having displacements of several feet [*Slemmons*, 1957]. The region is advantageous for study because it is crossed by a first-order triangulation net and a first-order line of levels, and also because the 1954 movements have been accurately measured by re-surveying [*Whitten*, 1957]. But the 1954 movements are only the latest of a long series of displacements during the latter part of the Cenozoic era, and gravity data help to determine the total displacement by indicating the thickness of light-weight fill in the basins. Structural movements in the Basin and Range Province are of special interest because the region is broken up by normal faults and, unlike many mountain regions, appears to be undergoing distension.

Gravity data also give an indication of the degree of regional and local isostatic balance, a matter of great interest because of its bearing on the problem of how isostatic compensation is

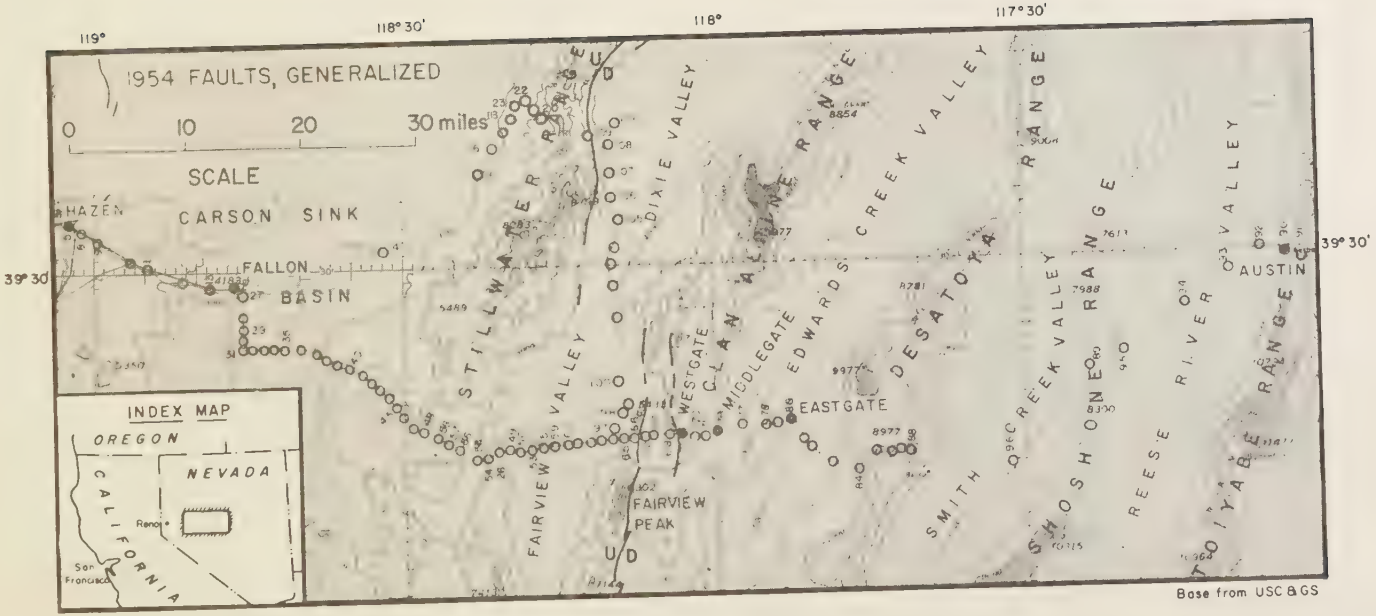


FIG. 1—Map of gravity stations, Hazen to Austin, Nevada

maintained as a broad region is raised to a high elevation. Another question that was investigated is whether a measurable change in gravity took place as a result of sudden shifting of masses during the faulting in 1954.

Gravity measurements—Gravity measurements were made in the summer of 1956 with Stanford's North American gravimeter operating with a sensitivity of 0.1263 mgal per scale division. The primary base station for all readings is the pendulum station of the U. S. Coast and Geodetic Survey at Mystic, California [*Duerksen*, 1949]. Another pendulum station, at Truckee, California, was tied to the Mystic station by gravimeter readings (Station 75, appendix). The pendulum determinations are stated only to the nearest mgal, but relative measurements by gravimeter are made to the hundredth of a mgal; hence the base value at Mystic, which is given as 979,630 mgal, was assumed to be 979,630.00 mgal.

The Bouguer anomaly as used in this paper is determined (in mgal) as:

g_{My}	+	g_d	+	$0.05998\,h$	−	g_0
Gravity at Mystic, US C & GS		Observed difference from Mystic		Elevation and simple Bouguer corrections; h is altitude in ft		Theoretical gravity at sea level

Combined corrections for tidal variation and instrument drift were made by returning to a previously occupied station about every two hours during each day. These corrections, which are incorporated in the 'observed difference from Mystic,' amount to roughly the expected tidal variation; the instrument itself was very stable. Terrain corrections were not made because no topographic maps are available for part of the region surveyed. Terrain corrections to a radius of 13.6 mi made by C. H. Sandberg in a similar region in western Nevada amounted to only about 2 mgal at most of the stations and a maximum of 10 mgal at one station [*Thompson* and *Sandberg*, 1958].

The precision of the field measurements is estimated to be within 0.1 mgal, assuming that the calibration constant of the instrument is correct within 0.1 pct. The calibration constant, originally determined by the Stanolind Oil and Gas Company, Tulsa, Oklahoma, was checked against other instruments, against the difference in observed gravity between pendulum stations, and against the change with elevation in a tall building; these measurements all agree in showing that the calibration constant is correct to better than 0.2 pct. Elevations of nearly all of the stations are known within 0.1 ft because the stations are at bench marks; hence elevation corrections have a precision of 0.01 mgal. A density of 2.67 g/cc was used in making the Bouguer correction, following the standard practice of the Coast and Geodetic Survey. The actual average density of pre-Cenozoic rocks that crop out in the mountains is probably within the range 2.6–2.8 g/cc. *Birch* [1950, p. 608] found an average of 2.67 in the Front Range of Colorado, and *Thompson* and *Sandberg* [1958] found densities within the stated range in western Nevada. The Cenozoic rocks, on the other hand, are generally much less dense.

Most of the gravity determinations were made with the instrument set up either directly over, or within a few feet of, permanent bench marks. Where the station was not directly over a bench mark, the elevation was determined from the bench mark by means of a hand level. Station locations are shown in Figure 1 and data are tabulated in the appendix.

Regional gravity and isostasy—The Bouguer anomaly averages −160 mgal in the western part of the area (Fig. 2) and −215 mgal in the eastern part. The change in anomaly from west to east, −55 mgal, indicates that an added slab of rock of density 2.67 g/cc and 1600 ft thick can be supported isostatically. The actual increase in average altitude over the same distance is estimated from topographic maps to be 1700 ft, which is in reasonably close agreement. The broad variations in Bouguer anomaly are thus related to isostatic compensation. Further support for this conclusion is provided by reconnaissance measurements which were extended westward from the area of Figures 1 and 2 to connect with a previously studied area near the Sierra Nevada, [*Thompson* and *Sandberg*, 1958]. In the regional picture that emerges, the basin of Carson Sink forms a gravity high and topographic low; westward toward the Sierra and eastward toward the higher Basin Ranges, the Bouguer anomaly grows more negative as the average altitude increases. This inverse relation between the Bouguer anomaly and the altitude of the land is the normal condition to be expected wherever compensation exists.

Although the pattern of the anomalies indi-

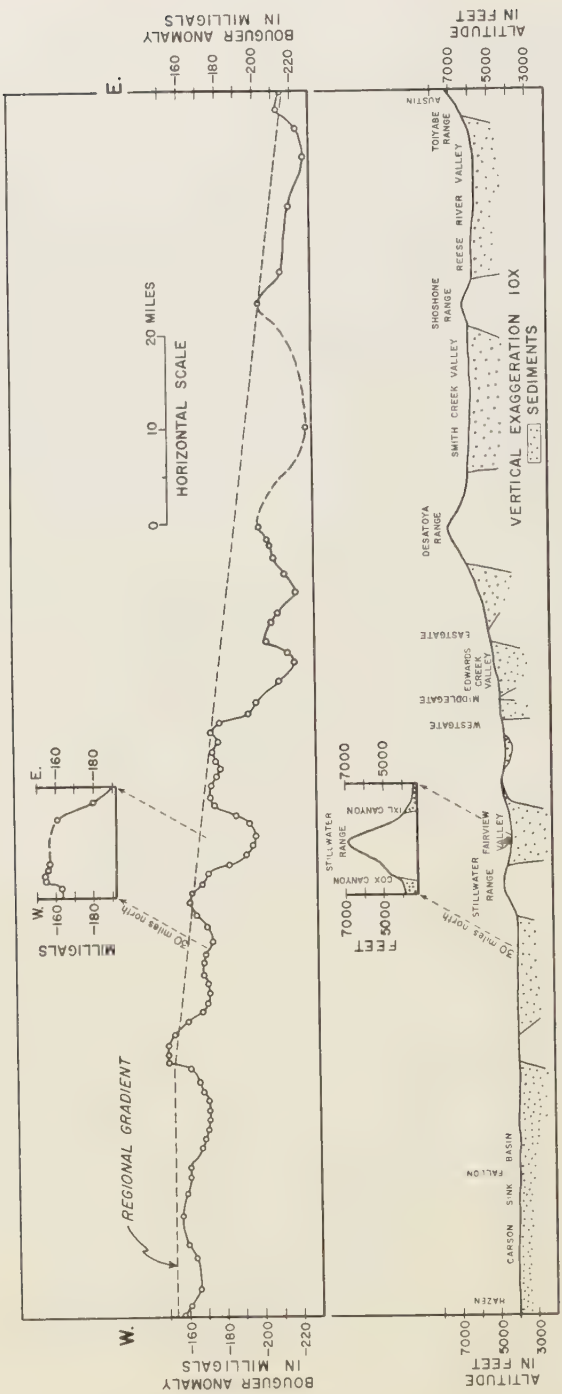


FIG. 2—Gravity profile and section, Hazen to Austin, Nevada

cates an approximation to regional isostatic balance, the average Bouguer anomaly appears to be roughly 10 pct larger than required for exact compensation. Stated in another way, the free-air anomaly averages about −20 mgal and is positive only at a few stations in the higher parts of the mountains. Part of this discrepancy may be attributed to a lack of terrain corrections and part to the (negative) attraction of compensating material of low density under the higher mountains lying east and west of the area, but negative isostatic anomalies at the pendulum stations in Nevada [*Duerksen*, 1949] suggest that these effects are not sufficient to explain all the discrepancy. Over-compensation sufficient to support topography a few hundred feet higher throughout the region is one possible interpretation. In any case the departure from regional compensation is small; compensation can safely be estimated at 100 pct ± 10 pct.

Apart from the degree of compensation is the interesting problem of the depth and form of the compensating material. Part of the compensation in this region, perhaps 10 pct to 15 pct of the total, must be in the uppermost part of the crust. That conclusion follows from the large amount of low density material in the basins (and parts of the mountains) as discussed in this paper. Probably more than half of the material above sea level is sedimentary debris and volcanic debris of low density. Assuming that half the material above sea level has an average density of 2.2 g/cc (Table 1) and the other half is rock of average density 2.7 g/cc, the average of all the material above sea level is approximately 2.4 g/cc. The difference between this average and 2.67 g/cc, the value assumed in computing the Bouguer anomaly, is 0.3 g/cc, an amount sufficient to account for 10 pct of the Bouguer anomaly. But the low-density material also extends below sea level in some of the basins, and it may possibly account for as much as 15 pct of the anomaly, on the average. Viewed another way, the upper part of the crust is less dense than usual and thus requires less compensating material at depth to support a given surface altitude.

The remainder (major part) of the compensation is probably not entirely effected by unusually light rocks within the crust, because this would require a density contrast of nearly 0.2 g/cc throughout the entire thickness of the crust. Variation in crustal thickness is generally thought to account for the compensation, but variations within the mantle are an additional possibility [*Tatel* and *Tuve*, 1955, p. 50].

Compensation and structure of basin ranges—No evidence of local compensation of the ranges can be found in the gravity data (Fig. 2). Ranges should show local negative Bouguer anomalies if they are compensated. ('Local anomaly' as used here refers to the anomaly in the ranges relative to that in the adjacent lowlands; the meaning is identical with the residual anomaly of exploration geophysics except that the residual anomaly is usually computed for much smaller areas.) In contrast to the regional picture, in which the Bouguer anomaly varies inversely with altitude and indicates compensation, the anomaly varies locally directly with altitude. It may be argued that use of a larger density than 2.67 g/cc for computing the Bouguer anomaly in the ranges would make the local anomalies negative. In order to make them negative, however, an improbably large density must be assumed. For example, on the west side of the Desatoya Range, where the anomaly varies directly with altitude, a density greater than 3.3 g/cc would be required in order to reverse the gradient of the anomaly, but such a large density is absurd. The rocks exposed in the range are light-weight rhyolite tuffs, and the anomaly is interpreted as indicating thinning of the tuffs toward the axis of the range.

Elsewhere in western Nevada, parts of the basins are free of sediments and a direct comparison of Bouguer anomalies in the ranges with those in sediment-free basins also fails to show local compensation [*Thompson* and *Sandberg*, 1958].

A lack of local compensation rules out any hypothesis of structural history that requires the formation of compensating roots beneath the ranges. For instance, a hypothesis of compressional folding and local thickening of the crust, followed by isostatic uplift of the thickened parts into ranges, is not in accord with the gravity picture. Four models that do accord with the gravity data are shown diagrammatically in Figure 3, which is not intended to portray the real structure but only to help define the permissible limits of interpretation of the gravity data. In (a) of Figure 3 the whole crust is shown as if warped by compressive deformation; regional compensa-

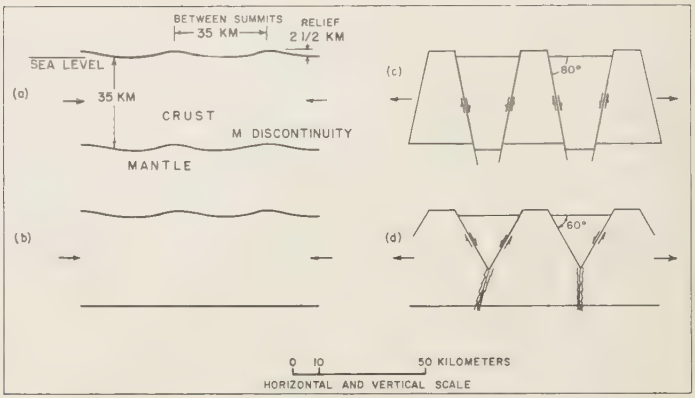


FIG. 3—Simplified models permitted by gravity data; horizontal arrows represent compression, (a) and (b) or extension, (c) and (d)

tion may be maintained, but the individual ranges are supported by the strength of the crust and are not locally compensated. The slight downwarp, (root) at the base of the crust below each basin is too small and too far below the surface to be readily detected in the gravity measurements; therefore the structure shown in Figure 3(b), in which only the upper part of the crust is deformed, is a reasonable alternative. In (c) of Figure 3 the crust is shown as if broken up by normal faults resulting from extension (with or without strike-slip components of movement). The mountain blocks have risen toward buoyant equilibrium and the valley blocks have sunk; they approach what *Vening Meinesz* [1948] has called crustal shear isostatic compensation. The 'roots' are again beneath the basins but are too small and too deep to be detected gravimetrically (except in larger structures than these, in areas uncomplicated by near-surface variations). Consequently, the structure illustrated by Figure 3(d), in which only the upper part of the crust is faulted, is an acceptable alternative. In Figure 3(d) the lower part of the crust is assumed to be extended plastically or to be dilated by igneous intrusions. Inasmuch as warped or tilted strata and normal faults with dips of about 60° are known features of the region, Figure 3(d) with the addition of shallow tilting and warping may best represent the actual structure.

Concealed structure of the basins—Local negative anomalies are associated with all the basins (Fig. 2). The anomaly relief is about 30 mgal at Fairview Valley and 40 mgal near Westgate. Gradients are so steep that the anomalous masses must be shallow; they cannot, for instance, li at the base of the earth's crust. The maximun gradient is about 20 mgal/mi; according to th formulation of *Bott* and *Smith* [1958, p. 3], th maximum depth to the top of the anomalou mass would be about one mile. These loca anomalies are interpreted as caused by low density fill of sedimentary and volcanic materia in the basins.

TABLE 1—*Density of rocks*

Rock	Measured range in density (Water-saturated rock
Sedimentary rocks, Cenozoic	1.7–2.5 g/cc
Volcanic rocks, Cenozoic	
Pyroclastic rocks	1.8–2.6
Flows	2.2–2.8
Granite rocks, Mesozoic and Cenozoic	2.62–2.67
Metamorphic rocks, Mesozoic	2.7–2.8

Table 1 shows a summary of about 50 labora tory measurements of density made on rock similar to those in the area under study [*Thomp son* and *Sandberg*, 1958]. The important poin is the large contrast between Cenozoic sed mentary and volcanic rocks and the olde rocks. The maximum difference is approximatel 1 g/cc and the average about 0.5 g/cc. Thi figure is rough but is certainly correct within factor of two.

In Figure 2 a regional gradient is drawn b connecting gravity values of bedrock stations i

the Stillwater Range, in the Toiyabe Range, and at Westgate; the other ranges are covered by Cenozoic volcanic rocks, which with minor exceptions have lower density than the pre-Cenozoic bedrock. The western end of the regional gradient is projected on the basis of bedrock stations that lie west of the area shown on Figure 2.

Using 0.5 g/cc as the most probable density contrast, one can readily compute the approximate thickness of fill in the basins. Carson Sink Basin (west side of Figs. 1 and 2) shows a local negative anomaly of 3 to 18 mgal. Thus the basin, although very wide, contains a maximum of only about 3000 ft of sediments along the line of the survey. West of Fallon the sediments appear to be only a few hundred feet thick. A spur of the Stillwater Range 12 mi east of Fallon bounds the basin. The spur is covered by basalt and the abrupt positive anomaly associated with it suggests that the basalt is not a single thin flow but is hundreds of feet thick. Alternatively, thin basalt may be underlain by bedrock of high density. Between the spur and the Stillwater Range is another basin, and the gravity data in that basin indicate a thickness of sediments of about 2000 ft.

The local gravity anomaly on the eastern side of the Stillwater Range is two to three times that on the western side, indicating a great thickness of low density material to the east, in Fairview Valley. These relations hold, not only on the main profile of Figure 2, but also on the short profile taken across the range 30 mi farther north (see inset above the Stillwater Range).

Fairview Valley is of special interest because of the 1954 fault movements (Fig. 1). Figure 4 shows the gravity measurements across Fairview Valley; the plotted points are the same as in Figure 2. Below the gravity profile in Figure 4 is a geologic section that is capable of explaining the gravity anomaly. The curve computed from the section and corrected for the regional gradient is shown by the solid line. The shape of the basin floor in cross section may be interpreted as representing step-faulting or alternatively as representing warping, as shown here. Or if we assume that the density contrast is smaller near the edges of the basin—as it is on the surface today because the center is occupied by low-density playa silts and the margins by gravels—the gravity data permit a simple flat-bottomed graben. The outstanding thing about

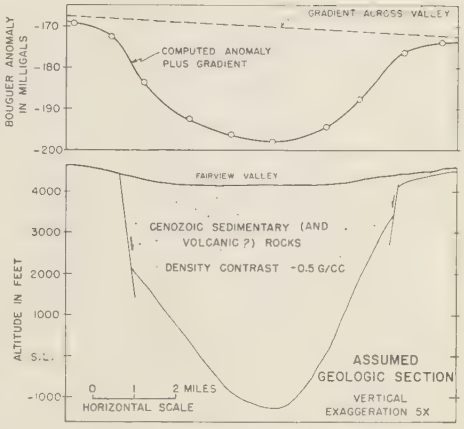


FIG. 4—Gravity profile and section across Fairview Valley

this section is the depth of the fill, about 5000 ft. The highest peaks in the adjacent ranges rise roughly another 5000 ft above the basin, making a total structural relief of 2 mi in a basin only 7 mi wide. We shall return to a discussion of this basin in connection with the fault movements.

From Fairview Valley to Westgate, where the gravity line crosses the 1954 faults, the anomaly indicates no more than a few hundred feet of sediments; the shallow bedrock in that area is part of the buried northern extension of Fairview Peak and of the southern extension of the Clan Alpine Range. Beyond Westgate the anomaly curve drops sharply and is interpreted as indicating a fault or faults with a downthrow to the east of more than 2000 ft. Volcanic rocks crop out at Middlegate but they have only slight influence on the gravity. In this area *Axelrod* [1956, p. 182–188 and Fig. 10] describes a 3000-ft section of silicic volcanic rocks of Cenozoic age overlain by a 3500-ft sedimentary section, the whole dipping 10° to 40° eastward. His description (p. 184) seems to permit the alternative interpretation of major intertonguing of the volcanic and sedimentary rocks, with the former mainly in the mountains and the latter mainly in the basins. The total thickness, 6500 ft, is in good agreement with the local gravity anomaly, which amounts to −40 mgal, and with the assumption of a negative density contrast of 0.5 g/cc for both the sediments and the volcanics. Axelrod interprets the west side of Middlegate Hills as bounded by a major fault with downthrow to the west, but the gravity data indicate

either that the fault is small or that there is little density contrast between the volcanic and the sedimentary rocks. On the other hand, the gravity data suggest a large fault with downthrow to the west at Eastgate, where volcanic rocks are again brought to the surface.

Between the Desatoya and Shoshone Ranges, both of which are covered by volcanic rocks of low density, Smith Creek Valley shows a local negative anomaly that suggests about 2000 ft of sedimentary material in the valley. The gravity high near the axis of the Desatoya Range probably indicates thinning of the volcanic cover toward the axis. The Reese River Valley contains an estimated 1500 to 3000 ft of sediments, and a fault boundary is suggested at the east side by the fairly steep gravity gradient.

Granitic bedrock crops out in the Toiyabe Range near Austin. In the whole distance between Westgate and Austin, low-density volcanic rocks and sediments are the only materials exposed, and from the positions of the contacts relative to the gravity anomaly, Cenozoic sediments and volcanics are estimated to contribute about equally to the local negative anomalies.

Gravity before and after 1954 faulting—The possibility of gravity changes accompanying the 1954 faulting remains to be considered. The 1954 faults generally follow the borders of the large basins, but where the main gravity profile crosses them, the faults depart from the usual habit and follow the margins of the small basin between Fairview Valley and Westgate. The west side of the small basin dropped a maximum of about 5 ft [*Whitten*, 1957, p. 324]. A movement of that magnitude, assuming that it is accompanied by no density changes in underlying rocks, should cause an increase of observed gravity amounting to the sum of the free-air and Bouguer corrections for an elevation difference of 5 ft. This increase should be 0.3 mgal. On the other hand, the Bouguer anomalies before and after faulting, each reduced to the elevation at the time of measurement, should show a difference only if masses have been redistributed or densities changed below the surface. Comparison of precise surveys before and after faulting may thus yield information on such changes.

The Standard Oil Company of California furnished data on a survey made previous to the 1954 faulting. Two of their base stations were reoccupied during the present survey and all data were reduced to the same base. The Bouguer anomalies from the Stillwater Range to Westgate were then compared. The largest uncertainty is in the location of the earlier stations, which, excepting the base stations, were not at permanent landmarks. In spite of this uncertainty, the two sets of data agree within 0.5 mgal. In the downfaulted block, two stations show Bouguer anomalies 0.2 and 0.5 mgal more negative after the faulting, and other stations agree within 0.1 mgal. As the uncertainty is also about 0.5 mgal, we can only conclude that the Bouguer anomaly either did not change or else decreased algebraically by an amount no greater than 1.0 mgal. If the anomaly in fact decreased, the most likely of several possible explanations is a slight decrease in density of shallow material in the disturbed zone.

Distension accompanying 1954 and earlier deformation—The total structural relief at Dixie and Fairview Valleys is roughly the sum of the topographic relief, about a mile, and the buried relief, another mile, making a total of two miles. Large horizontal distension, normal to the trend of the valleys, is also required. If the basin is bounded by faults dipping 60° (the angle shown in the section, Fig. 4), without major warping of the strata, the extension normal to the strike amounts to about a mile on each side of the basin, a total extension of two miles. If the faults dip 70°, the extension amounts to about a mile and a half.

Of this large deformation the last part, which took place in 1954, is known most accurately because of seismic and geodetic measurements. Figure 5 summarizes the horizontal displacements determined by the U. S. Coast and Geodetic Survey from triangulation before and after the

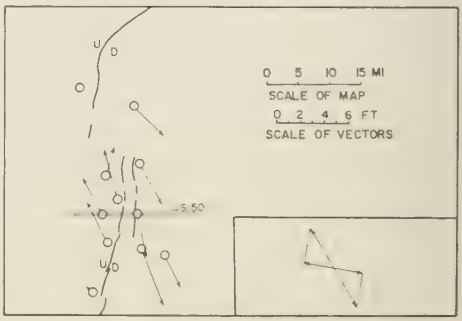


Fig. 5—Horizontal movements close to the faults (after Whitten)

1954 faulting [*Whitten*, 1957]. Epicenters of the two large earthquakes, which occurred in succession four minutes apart, were determined by *Romney* [1957] to lie about 35 mi from each other, one near the northern and one near the southern edge of Figure 5. He estimated the focal depths from the time interval $pP - P$ to be 15 km for the southern earthquake and 40 km for the northern. In Figure 5 the vectors indicate displacement of triangulation points relative to points about 40 mi east and west of the faults, which had no displacement relative to each other. On the average, from any point at the fault, points on opposite sides moved away to the northwest and southeast (Fig. 5, inset). The component of extension normal to the fault trace (horizontal component of dip slip) is 5 ft.

Dip of the 1954 faults can be determined in three ways, and the agreement among the results verifies the measurements of extension. First, the vertical and horizontal components of dip slip can be used to determine dip. The vertical displacement is 7 ft in Dixie Valley [*Whitten*, 1957] and is also 7 ft where the fault is in bedrock east of Fairview Peak [*Slemmons*, 1957, p. 360]; where U. S. Highway 50 crosses the fault zone the sum of displacements on oppositely facing faults is also 7 ft [*Whitten*, 1957, p. 324]. If 7 ft is taken as an average vertical displacement and 5 ft as the average extension, tangent of the dip is 7/5 and the dip is 55°. A second measure of the dip is provided by seismic first-motion observations. By this method *Romney* [1957] determined a dip of 62° at the southern focus. The third measure, direct observation, shows dips in bedrock of 55 to 75° [*Slemmons*, 1957]. Agreement among these three determinations of dip is reasonably good.

The close correspondence of dip and direction of motion at the surface with the same quantities determined at the focus of the first earthquake [*Romney*, 1957] is of special importance to an understanding of the mechanics of deformation. The focal depth of 15 km, coupled with the agreement just mentioned, implies that the fault or fault zone extends continuously from the surface to a depth greater than 15 km and that therefore a large fraction of the total thickness of the earth's crust is being deformed in the same manner. The probable depth of the second earthquake was even greater, 40 km [*Romney*, 1957]. Faulting observed at the surface would therefore seem to represent the type of deformation prevailing throughout most of the thickness of the earth's crust; the whole crust is undergoing extension. Of the four diagrams in Figure 3, 3(d) would come closest to fitting the facts.

One note of caution must be mentioned in regard to the focal depths. Romney's data on the epicentral locations, focal depths, and fault dips are not entirely self-consistent. The two epicenters lie close to the surface trace of the fault, whereas they should lie 9 and 23 km east of the trace if the foci are 15 and 40 km deep and if the fault dips 60°. Since the epicentral locations are not likely to be in error by as much as 9 and 23 km, the discrepancy casts doubt either on the focal depths or else on the continuity of the fault zone at a dip of 60°. The easiest escape from this dilemma is to assume that the phase that Romney tentatively identified as pP was in reality some other phase and that the foci were near the surface. Alternatively, the fault zone might be nearly vertical, although individual faults at the surface and at the focus dip 60° eastward.

A crude test of the depth of deformation can be made by analyzing the volume of the structural depression as related to the amount of extension. The cross-sectional area of the 1954 depression at U. S. Highway 50 [*Whitten*, 1957, p. 324] is approximately 6000 m^2. Uniform extension of 1.5 m (5 ft) to a depth of only 4 km would create a void large enough to account for the depression; if the potential void was triangular in cross-section, with a width of 1.5 m at the surface tapering downward to zero, the total depth would still be only 8 km. A similar calculation for the total deformation can be made as follows. The cross-sectional area of the fill in Figure 4 is approximately 13 km^2 and this amount must be at least doubled to include the relief between the surface of the fill and the mountain tops. On this basis the total extension of 2.4 km (1.5 mi) would require a triangular void roughly 20 km deep. None of these estimates are nearly as large as the larger of the two focal depths in question.

Rate of distension in the Basin and Range Province—The data indicate that the region of Dixie and Fairview Valleys has been distended in a nearly east-west direction about a mile and a half. If we assume that each of the principal basins between the Sierra Nevada and the Wasatch Mountains has been deformed this much on the average, the total distension amounts

TABLE 2—*Historic faults*

Location	Date	Estimated east-west component of extension (ft); dip, if unknown, is assumed to be 60°	Reference
Owens Valley, Calif.	1872	13	Hobbs [1910]
Pleasant Valley, Nev.	1915	13	Page [1935]
Cedar Mountain, Nev.	1932		Gianella and Callaghan [1934]
Excelsior Mountains, Nev.	1934	0.1	Callaghan and Gianella [1935]
Fort Sage Mountains, Calif.	1950	0.4	Gianella [1957]
Fallon, Nev.	1954	1	Tocher [1956]
Dixie Valley-Fairview Peak, Nev.	1954	5	Whitten [1957]

to 30 mi or 5 pct. And if the deformation took place in the last 15 million years, as suggested by the geologic history (deformation of Miocene-Pliocene and younger rocks), the rate is 2 mi/million years or only 1 ft/century. The rate of extension indicated by several fault movements within historic times appears to be at least 1 ft/century (Table 2). The faults listed in Table 2 lie in a north-south belt about 250 mi long. For at least this distance the data are consistent with an extension of 1 ft or more in the last hundred years. Prehistoric Quaternary faults are also numerous; they strongly suggest that the historic rate of deformation is not abnormally high.

It is recognized that tilted and warped blocks, which are a characteristic feature of the region, may absorb all or a part of the extension that occurs by faulting. No evidence for this view can be found in the geodetic measurements across the 1954 fault zone, however, although the valley block was tilted at that time. As *Whitten* [1957, p. 323] lucidly stated, blocks on opposite sides of the fault zone "are not only shearing but also pulling apart, creating a void that explains the lowering of the Valley adjacent to these faults." Probably expansion of tilted blocks by internal disruption along many minor faults is sufficient to balance any apparent contraction due to the tilt. The geodetic measurements are consistent with the elastic rebound theory of faulting, and we must conclude that a large region is undergoing slow distension, which from time to time results in breaks like those of 1954.

Strike-slip displacement—Throughout this paper, the strike-slip component of fault displacement has been neglected, for the interpretations presented are entirely independent of the strike-slip component. Strike-slip movement does not change the area of any given region, but any component of dip slip on non-vertical faults must result in a change of area. This fact is of cardinal importance in the search for basic causes and processes. If the area is expanding, as it seems to be, the often expressed idea that basin-range structure is the superficial result of horizontal compression in the deeper crust can hardly be correct.

It is not yet clear how fundamental the strike-slip component of the 1954 faults may be. Any normal fault that varies in strike should have a strike-slip component of displacement on some of its segments. In agreement with this, the southern part of the 1954 fault zone shows clear evidence of a large right-lateral component, but the northern part exhibits almost pure dip slip. The geodetic measurements also indicate that uniform displacement of the blocks east and west of the faults would produce little or no strike slip on the northern part of the fault, which trends northeast (Fig. 5). Throughout their extent the fault surfaces have a dip characteristic of normal faults, rather than the vertical dip to be expected of strike-slip faults on mechnical principles [*Anderson*, 1951; *Hubbert*, 1951].

It is also interesting to note that the seismically determined direction of displacement at the second epicenter should have been dip slip to agree with the surface evidence but could not be established because motion was still continuing from the first earthquake.

Strike-slip displacement, if it has a significant pattern in the Basin and Range Province, is an

additional important key to the mechanics of the deformation. In the Dixie-Fairview region, the geodetic displacement vectors lie along a line that is not normal to the average trend of the basins and ranges; an average right-lateral component of displacement is indicated. If the geodetic measurements are representative of deformation in a larger part of the Basin and Range Province, the entire region is undergoing right lateral deformation in addition to distension.

Summary of conclusions—Gravity measurements show that the region is approximately in isostatic balance, but the possibility of slight overcompensation is not excluded. Sedimentary and volcanic material that occupies a large volume in the basins and in parts of the mountains has a low density and accounts for roughly 10 to 15 pct of the average Bouguer anomaly. Consequently the compensation for this high region is partly in the superficial materials themselves. Although the area is regionally compensated, the gravity data provide no evidence for compensation of individual basins or ranges.

Large negative anomalies are associated with all the basins and indicate great thicknesses of sedimentary and volcanic fill. Fairview and Dixie Valleys contain more than 5000 ft of fill and Edwards Creek Valley more than 6000 ft.

The Bouguer anomaly in the area downfaulted in 1954 either did not change during the faulting or else decreased algebraically by an amount no greater than 1.0 mgal.

Distension of 5 ft in 1954, measured in a direction normal to the strike of the faults, is the latest increment of a total distension that amounts to about $1\frac{1}{2}$ mi in 15 million years. If this small region is a fair sample of the Basin and Range Province, the total distension in the province is 30 mi in 15 million years and the rate of distension is about 1 ft/century, an amount well within the rate indicated by historical fault movements.

The analysis presented is independent of a strike-slip component of displacement, which occurred on only part of the fault zone, but a component of strike slip along the average trend of the ranges is an added factor. A hypothesis of deep-seated horizontal compression as a mechanism of origin of basin-range structures is not compatible with the implied expansion of area, nor does strike-slip displacement explain the expansion.

Acknowledgments—The gravity measurements were made during the tenure of a National Science Foundation postdoctoral fellowship at Lamont Geological Observatory. To his colleagues at Lamont the writer is indebted for numerous courtesies and many valuable discussions of geophysical problems. Ben M. Page and Konrad B. Krauskopf, of Stanford University, made helpful suggestions, and Page generously contributed geologic data. C. H. Sandberg, of the U. S. Geological Survey, supplied information on gravity tie points. The Standard Oil Company of California, through the courtesy of William Barbat and E. G. Dobrick, Jr., furnished gravity measurements made prior to the 1954 faulting.

References

Anderson, E. M., *The dynamics of faulting*, second edition, 206 pp., Oliver & Boyd, London, 1951.

Axelrod, D. I., Mio-Pliocene floras from west-central Nevada, *Calif. Univ. Pubs. Geol. Sci.*, **33**, 1–322, 1956.

Birch, F., Flow of heat in the Front Range, Colorado, *Bull. Geol. Soc. Amer.*, **61**, 567–630, 1950.

Bott, M. H. P., and R. A. Smith, The estimation of the limiting depth of gravitating bodies, *Geophys. Prospecting*, **6**, 1–10, 1958.

Callaghan, E., and V. P. Gianella, The earthquake of January 30, 1934, at Excelsior Mountains, Nevada, *Bull. Seism. Soc. Amer.*, **25**, 161–168, 1935.

Duerksen, J. A., Pendulum gravity data in the United States, *U. S. Coast and Geodetic Survey Special Publication* 244, 218 pp., 1949.

Gianella, V. P., Earthquake and faulting, Fort Sage Mountains, California, December, 1950, *Bull. Seism. Soc. Amer.*, **47**, 173–177, 1957.

Gianella, V. P., and E. Callaghan, The Cedar Mountain, Nevada, earthquake of December 20, 1932, *Bull. Seism. Soc. Amer.*, **24**, 345–384, 1934.

Hobbs, W. H., The earthquake of 1872 in the Owens Valley, California, *Beitr. Geophysik*, **10**, 352–385, 1910.

Hubbert, M. K., Mechanical basis for certain familiar geologic structures, *Bull. Geol. Soc. Amer.*, **62**, 355–372, 1951.

Page, B. M., Basin-Range faulting of 1915 in Pleasant Valley, Nevada, *J. Geol.*, **43**, 690–707, 1935.

Romney, C., The Dixie Valley-Fairview Peak, Nevada, earthquakes of December 16, 1954: Seismic waves, *Bull. Seism. Soc. Amer.*, **47**, 301–319, 1957.

Slemmons, D. B., The Dixie Valley-Fairview Peak, Nevada, earthquakes of December 16, 1954: geological effects, *Bull. Seism. Soc. Amer.*, **47**, 353–375, 1957.

Tatel, H. E., and M. E. Tuve, Seismic exploration of a continental crust, in Poldervaart, A., ed., Crust of the earth—a symposium, *Geol. Soc. Amer. Special Paper 62*, pp. 35–50, 1955.

Thompson, G. A., and C. H. Sandberg, Structural significance of gravity surveys in the Virginia City-Mount Rose area, Nevada and California, *Bull. Geol. Soc. Amer.*, **69**, 1269–1282, 1958.

Tocher, D., Movement on the Rainbow Mountain fault, *Bull. Seism. Soc. Amer.*, **46**, 10–14, 1956.

Vening Meinesz, F. A., Gravity expeditions at sea, 1923–1938, *Netherlands Geodetic Comm. Pub.*, **4**, 1948.

Whitten, C. A., The Dixie Valley-Fairview Peak, Nevada, earthquake of December 16, 1954; geodetic measurements, *Bull. Seism. Soc. Amer.*, **47**, 321–325, 1957.

(Manuscript received November 8, 1958; presented at the Thirty-Eighth Annual Meeting, Washington, D. C., April 30, 1957.)

Appendix

Gravity Stations between Hazen and Austin, Nevada, 1956.
Base is USC & GS pendulum station at Mystic, California.

Sta. No.	Bench Mark	Latitude North	Elevation of Instrument Feet	Observed Gravity Milligals	Int. Formula Gravity Milligals	Bouguer Anomaly Milligals
4	(Mystic)	39° 27.1′	5152.4	979, 630.00	980, 131.8	−192.8
5	B48	33.5	4002.0	744.80	141.3	−156.5
7	T47	28.75	3964.6	735.44	134.3	−161.1
8	A48	33.05	4005.2	740.05	140.7	−160.5
9	Z47	32.30	4031.2	732.07	139.6	−165.8
10	L321	30.86	4026.7	732.09	137.4	−163.8
11	F388	30.25	4017.4	736.07	136.5	−159.5
12	V47	29.45	4000.5	738.81	135.3	−156.6
13	U47*	28.82	3976.1	736.45	134.4	−159.5
14		31.15	3898.0	739.10	137.8	−164.9
15	F389	37.0	3939.3	748.82	146.5	−161.4
16	K389	38.7	3948.1	745.18	149.0	−168.1
17	L389	40.0	3954.4	754.12	150.9	−159.6
18	M389	40.9	3946.0	755.98	152.2	−159.6
19	R387	41.3	5014.2	695.66	152.8	−156.4
20	5288′	41.0	5282.0	678.89	152.4	−156.8
21	M387	41.7	4778.3	712.26	153.4	−154.6
22	P387	42.0	4210.7	746.48	153.9	−154.9
23	3978′	41.8	3970.7	752.33	153.6	−163.2
26	Y46	16.6	4610.2	670.76	116.4	−169.2
27	Q47	28.3	3956.8	735.38	133.7	−161.0
28	P47	26.8	3955.0	726.95	131.4	−167.3
29	A388	25.9	3950.2	724.18	130.1	−169.0
30	N47	25.0	3949.7	721.18	128.8	−170.8
31	B388	24.1	3941.4	719.61	127.4	−171.4
32	M47	24.1	3933.5	720.15	127.4	−171.4
33	J297	24.1	3928.9	721.49	127.4	−170.3
34	L47	24.1	3928.3	723.22	127.4	−168.6
35	Q383	24.1	3925.3	725.56	127.4	−166.4
36	R383	24.15	3925.6	730.40	127.5	−161.7
37	T383	23.5	3963.9	738.67	126.5	−150.1
38	H47	23.25	3938.7	739.81	126.2	−150.2
39	U383	22.8	3955.4	737.97	125.5	−150.3
40	3964′	22.47	3958.2	733.73	125.0	−153.9
41	3	21.9	3942.2	726.05	124.2	−161.7
42	F47	21.43	3924.7	719.99	123.5	−168.2
43	N383*	20.9	3923.3	716.15	122.8	−171.4
44	E47*	20.23	3922.5	714.29	121.7	−172.2
45	G383*	19.5	3916.6	714.15	120.6	−171.6
46	H383	18.9	3914.5	715.47	119.8	−169.6
47	V383*	18.0	3918.3	713.51	118.4	−169.9
48	X383*	17.8	3903.9	713.40	118.1	−170.6
49	Z382*	16.4	4459.5	676.13	116.1	−172.5
50	X46	16.3	4274.3	676.12	115.9	−183.5
51	W46	16.7	4154.5	671.27	116.5	−196.1
52	S46	17.2	4391.5	676.80	117.3	−177.1

APPENDIX—(Continued)

Sta. No.	Bench Mark	Latitude North	Elevation of Instrument Feet	Observed Gravity Milligals	Int. Formula Gravity Milligals	Bouguer Anomaly Milligals
53	Y382	39° 16.7	4167.0	979,674.19	980,116.5	−192.4
54	E383*	16.3	4570.2	677.79	115.9	−164.0
55	Z46	16.0	4375.2	690.62	115.5	−162.5
56	A47	16.1	4063.4	707.13	115.6	−165.8
57	B47	16.8	3804.7	711.27	116.7	−171.3
58	Y383*	17.3	3896.4	709.34	117.4	−174.4
59	T382*	16.9	4149.0	670.18	116.9	−197.9
60	V46	17.0	4188.5	671.57	117.0	−194.3
61	U382	17.0	4284.9	672.57	117.0	−187.5
62	4431′	17.0	4423.8	675.61	117.0	−176.1
63	U46 AZI*	17.1	4557.8	670.03	117.1	−173.7
64	V382*	17.1	4596.3	666.65	117.1	−174.8
65	T46	17.2	4564.5	666.18	117.3	−177.4
66	W382	17.2	4423.7	672.84	117.3	−179.2
67	X382*	17.3	4435.8	676.06	117.4	−175.3
68	R46	17.4	4469.7	671.16	117.6	−178.4
69	D383*	17.4	4532.7	671.13	117.6	−174.6
70	Q46	17.3	4590.0	663.06	117.4	−179.1
71	Q130	17.3	4623.5	645.69	117.4	−194.4
75	(Truckee)	19.6	5889.4	583.07	120.8	−184.3
76	P46	17.5	4667.4	639.52	117.7	−198.3
77	N46	18.0	4782.3	620.98	118.4	−210.6
78	M46	18.0	5017.5	598.45	118.4	−219.1
79	A383	18.5	5029.8	602.92	119.1	−214.5
80	L46	18.5	5110.5	608.29	119.1	−204.3
81	K46*	17.0	5333.6	589.94	117.0	−207.2
82	S388*	17.0	5410.3	581.64	117.0	−210.9
83	Q388	17.5	5726.0	554.36	117.7	−220.0
84	T388*	17.0	6100.6	536.79	117.0	−214.4
85	V388*	17.5	6585.5	514.03	117.7	−208.8
86	F385*	17.0	6968.0	492.41	117.0	−206.7
87	G385*	17.0	7187.3	480.56	117.0	−205.4
88	E385	16.0	7301.2	476.67	115.5	−201.0
89	A935X . .	22.0	6431.2	536.62	124.3	−202.0
90	R63	30.0	6573.7	529.02	136.1	−212.9
91	Q385*	29.5	7179.9	489.90	135.4	−214.9
92	B384	30.5	5998.3	554.32	136.8	−222.8
93	H45*	29.5	5739.7	564.83	135.4	−226.4
94	J384*	26.5	5743.3	567.67	130.9	−218.8
95	P45	23.0	5967.3	553.90	125.8	−214.0
96	X45	14.5	6096.7	521.23	113.3	−226.5
97	A386	17.9	4461.9	675.70	118.3	−175.0
98	C386	19.1	4342.5	682.23	120.0	−177.4
99	4250′	19.7	4243.1	688.76	120.9	−177.7
100	F386*	21.2	4180.6	700.32	123.2	−172.2
101	L386	25.9	3998.1	709.85	130.1	−180.5
102	P386	28.3	3900.3	716.48	133.6	−183.2
103	R386*	29.9	3831.8	722.78	136.0	−183.4
104	T386	31.4	3784.1	723.07	138.2	−188.2
105	V386*	33.3	3701.7	726.17	141.0	−193.8
106	X386	34.9′	3620.3	736.31	143.4	−190.0
107	2386*	36.7	3581.2	754.64	146.0	−176.6
108	D387*	38.7	3531.1	758.96	149.0	−178.3
109	E387*	39.6	4028.8	747.74	150.3	−161.0
110	B387	39° 40.4	3482.4	979,761.09	980,151.5	−181.6

* Station was a few feet from bench mark; elevation determined from bench mark by hand level.

Some Seismic Profiles Onshore and Offshore Long Island, New York

M. Blaik, J. Northrop, and C. S. Clay

*Columbia University, Hudson Laboratories, Dobbs Ferry, N. Y.**

Abstract—Seismic velocity determinations in shallow layers (less than 2000 ft deep) have been made recently at an experimental location on the Atlantic Coast. Two reversed refraction profiles were obtained offshore. An unreversed refraction profile was completed on the beach. Reflection profiles were obtained both on the barrier beach and in the adjoining lagoon. Speed layering indicated by seismic reflection and refraction is compared with borehole velocity measurements. Significant thin high-speed and low-speed layers are shown by the borehole velocity data but not by the seismic profile data. The calculations from reflection and refraction shooting show reasonable agreement with the results of the borehole velocity survey for thick layering.

Introduction

Seismic velocity profiles are generally accepted as good approximations to actual subsurface velocities in many areas where direct measurements are not available. The purpose of the present work is to test the validity of this assumption at an Atlantic Coast location on Long Island. At the site selected for observations, 1941 ft of unconsolidated sediments overlie the basement (Fig. 1). The velocity distribution was determined by velocity logging (Fig. 2). This was accomplished by exploding $\frac{1}{2}$-lb charges 3 ft down in sand near a borehole, and recording the outputs of geophones fixed at 40-ft intervals in the uncased borehole. Basement velocities of from 16,000 to 16,740 ft/sec were measured and were found to compare well with previous measurements in the area [*Ewing* and others, 1950; *Oliver* and *Drake*, 1951; *Carlson* and *Brown*, 1955]. The basement is overlain by the Upper Cretaceous Raritan and Magothy? formations mapped by *Suter* and others, [1949]. The Raritan consists of a basal sand (Lloyd sand member) and an overlying clay member of laminated silty and solid clays with subordinate sandy layers. The Magothy? formation consists of alternating layers of fine clayey sands, fine sands, silts, solid clays, and several coarse water-bearing zones of sand and gravel. The driller's log shows layers of lignite which correspond to the low velocity layers. Gamma ray, neutron log, self potential, and resistivity curves for the section are shown in Figure 2 along with the velocity log and generalized driller's log.

For the completion of a pair of reversed seismic refraction profiles offshore (Fig. 3), the standard seismic refraction technique was employed [*Ewing* and others, 1937]. The instrument system gave an essentially flat frequency response and an oscillograph camera was used to record traces for various filters indicated on Figure 4. Shot sizes ranged from $\frac{1}{2}$ lb to 3 lb; shot distances were determined from water wave travel times. Water velocities were determined from tables with data from bathythermograph lowerings and salinity measurements.

The recording system utilized on land gave a flat frequency response from 10 to 200 cps and was operated with fixed gain. All recordings were made with six vertical component geophones per trace. For reflection recording, single geophones were placed 10 ft apart in line, and each group of six was spaced 50 ft with a spread length of 450 ft. For refraction recording, each group of six geophones was placed in a small (2 ft^2) area with a group spacing of 100 ft for a total spread length of 900 ft. Shot holes were about three feet deep in loose sand. Reflection shots consisted of $2\frac{1}{2}$-lb charges in three holes spaced 25 ft in line. Refraction shots were made with 5-lb charges in three holes, with 25-ft spacing, perpendicular to the geophone arrays. All shots were fired electrically. Reflection recordings in a shallow lagoon adjacent to the experimental site were made with the same system used on land, but barium titanate hydrophones (six per trace, 10 ft apart, in line) with 60-ft group spacing and

* This work was supported by the Office of Naval Research. Hudson Laboratories Contribution No. 26.

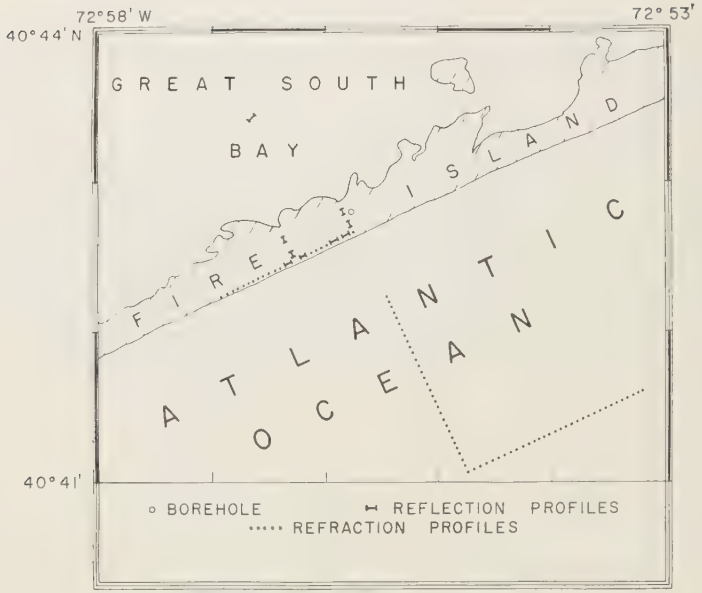


FIG. 1—Experimental Site, Fire Island, New York

240-ft spread length were substituted for the geophones. Shots were floated to avoid bubble pulse.

DATA AND INTERPRETATION

The data were fitted to the travel time curves by the method of least squares (Figs. 3 and 5). The slopes and intercepts thus obtained were used to calculate the velocities and thicknesses of layers. An evaluation of the experimental error involved in the calculations of thicknesses and velocities of the layers were obtained from Student's t tests using eighty per cent confidence limits. Thicknesses are believed to be accurate within about ten per cent, and velocities determined similarly have an experimental uncertainty of about five per cent.

Offshore refraction profiles—In order to determine the velocity of the first sedimentary layer for the marine profiles, shots were fired at very close range. Using these data combined with the zero intercept (bottomed shots and hydrophone) and later second arrivals, it was possible to determine a good travel time curve for the first sedimentary layer. A more reliable method of determining the velocity and thickness of the first sedimentary layer by CW signals has been discussed by *Tolstoy* [1958]. The second sedimentary layer was well determined by the first arrival on profiles 1 and 2, Figure 3.

The third sedimentary layer was defined by secondary arrivals; however, it was considered reliable. This interface correlates well with the Raritan-Magothy interface projected from Long Island and with a layer of similar velocity defined by first arrivals on a nearby profile (*Brown, Northrop, Frassetto,* and *Grabner,* in preparation). It also correlates with layers of similar velocity determined from second arrivals at nearby Shinnecock Inlet by *Oliver* and *Drake* [1951] and at Ambrose Lightship by *Carlson* and *Brown* [1955]. As there always is doubt about interpretations based on second arrivals, a preliminary analysis of the frequencies of the ground arrivals was made.* The arrivals that fall on the semiconsolidated V_3 line came in at about 20 cps, whereas the frequencies dominant on the overlying unconsolidated V_2 layer were 16–18 cps. It is interesting to note that the

* This analysis was done only on recordings for large shooting distances, that is, those recordings showing a considerable spread in time of the various ground arrivals.

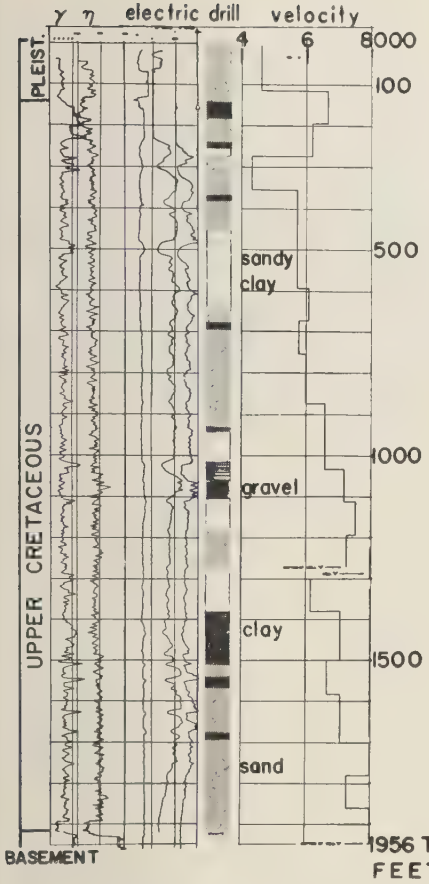


FIG. 2—Borehole Logs. These are, from left to right, gamma ray, neutron, spontaneous potential, resistivity with two electrode spacings, generalized driller's log, and interval velocity log

arrivals indicating the higher velocity, deeper interface appear to have a slightly higher frequency. A possible explanation for this may be that the separate arrivals may correspond to various group velocity maxima [*Tolstoy*, 1956].

The top surface of the basement dips slightly shoreward with a slope of 0.25°. This reversal of regional slope indicates that the depression found off Shinnecock Inlet by Oliver and Drake may be a characteristic feature of the basement along the south shore of Long Island or it may be due to local basement topography [*Beckman, Roberts* and *Luskin*, 1958]. However, this apparent reversal of regional dip may be due to the inability of the refraction method to show the presence of masked high speed layers. At the base of the sedimentary column, the velocity log (Fig. 2) shows a 7000-ft/sec layer down to the basement and the beach refraction profile shows a 540-ft layer with a velocity of 6990 ft/sec.

Onshore refraction profile—Arrivals T_1 through T_6 (Fig. 5) were included in the boundary refraction computations. Although this profile is unreversed, the velocities are believed to be sufficiently accurate approximations to compressional wave velocities in the layers represented. The justification of this view is strengthened by considering other local geological data [*Suter* and others, 1949] and seismic data [*Oliver* and *Drake*, 1951] which indicate that the above-mentioned profile is very nearly parallel to the general strike of the subsurface layers.

Among the arrivals identified and labeled on the graph of Figure 5a, three outstanding phases, *PSP*, R_1, and R_2, warrant particular attention. Interpretation of *PSP* as a boundary refraction wave traveling with Rayleigh wave velocity at the basement complex is consistent with the other refraction data in the same profile. A least-squares straight-line fit for these data gives $T = .489 + x/10{,}000$. Experiments and theory in the work of other investigators suggest that this is most likely to be a pseudo-Rayleigh wave [*Strick* and others, 1956; *McMillen*, 1946]. R_1 and R_2 are taken to be primary and secondary reflections from the sediment-basement interface.

In order to determine the compressional speed in the first layer, a short profile was made with single geophones spaced ten feet apart. The results have been reproduced as a travel time graph (Fig. 5b). From this it is clear that the dry beach sand layer is only a few feet thick and has a compressional wave speed of about 700 ft/sec or even less. Low values of compressional velocity of dry loose sand of this order of magnitude have been reported by numerous authors, including *Domzalski* [1956]. The short profile also shows an intermediate layer with a velocity of about 4400 ft/sec. The large profile was simplified by assuming an average velocity of 2000 ft/sec for the first layer.

Reflection profiles—Reflection recordings were taken of four reversed profiles on the beach (Fire Island, Fig. 1) and one reversed profile in the adjacent lagoon (Great South Bay, Fig. 1). The

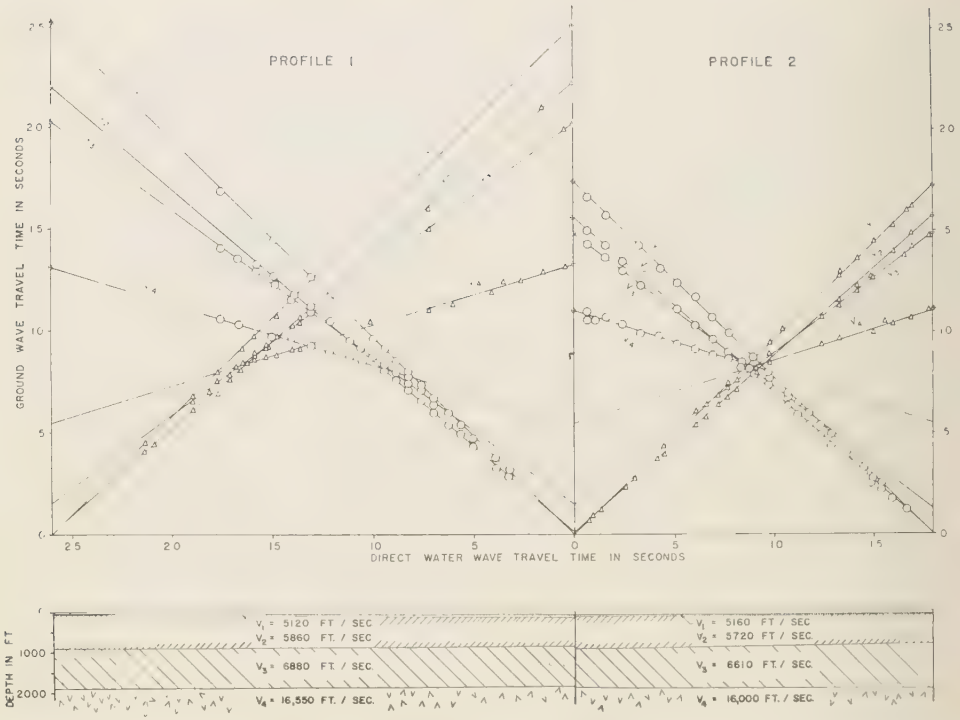


FIG. 3—Travel time curves for offshore pair of reversed refraction profiles. Geologic section on one-to-one scale shows small dips encountered and close agreement of profiles at intersections

INSTRUMENT	GAIN	FILTER
HYDROPHONE 1	HIGH	500 HIGH PASS
HYDROPHONE 1	LOW	500 HIGH PASS
HYDROPHONE 1	HIGH	5 - 30
HYDROPHONE 1	LOW	5 - 30
HYDROPHONE 1	HIGH	30 - 100
HYDROPHONE 1	INTERMEDIATE	30 - 100
HYDROPHONE 1	LOW	30 - 100
HYDROPHONE 1	HIGH	175 - 200
RADIO - TIME BREAK		
HYDROPHONE 2	HIGH	500 HIGH PASS
	OFF	
HYDROPHONE 2	HIGH	5 - 30

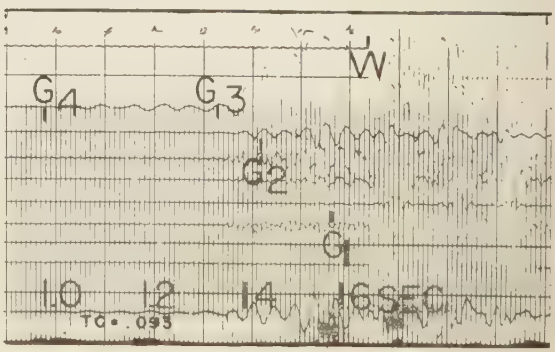


FIG. 4—Sample offshore refraction record; ground arrival subscripts correspond to velocity subscripts on Fig. 3

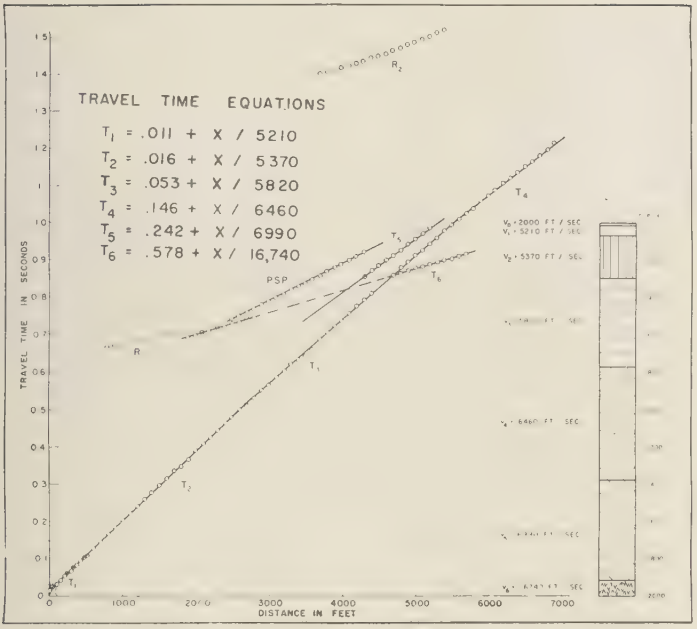


FIG. 5a—Onshore refraction profile, travel time curves and velocity formulas

lagoon profile was located about one mile northwest of the Coast Guard dock in eight feet of water. The beach recordings are particularly clear and have been reproduced together with a plot of profile positions in Figure 6. The reflections indicated on the recordings are believed to correlate between profiles almost a mile apart. Point plotted reflection depth profiles were constructed by use of wave front charts based on a linear velocity function fitted to the interval velocity data of Figure 2. A beach profile with the lagoon profile for comparison is shown in Figure 7. In general, the lower frequency beach profiles had many more coherent reflections than the lagoon profile, but in the latter case the reflection plots show less scatter. Apparently more high frequency noise was present in shallow water recordings. This effect, however, did not cause significant phase shifts in the reflections that were coherent.

Average and interval velocities were computed for all reflections which appeared coherent and correlatable on both halves of each reversed profile. The method of *Dix* [1955] was used in a modified form as follows: Let T_A and T_B designate reflection travel times from shotpoints A and B, respectively, to a detector between them. Also, let x be the distance from shotpoint A to the detector and d be the distance from shotpoint A to shotpoint B. Then following *Dix* [1955, p. 78] and taking the difference $T_A^2 - T_B^2$ gives

$$T_A^2 - T_B^2 = x \frac{2d + 4d \sin^2 \theta}{\bar{V}^2} + \frac{2VT_0 d \sin \theta - d^2}{\bar{V}^2}.$$

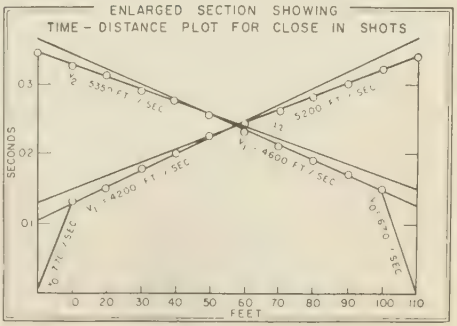


FIG. 5b—Enlarged section showing time-distance plot for close-in shots of onshore refraction profile

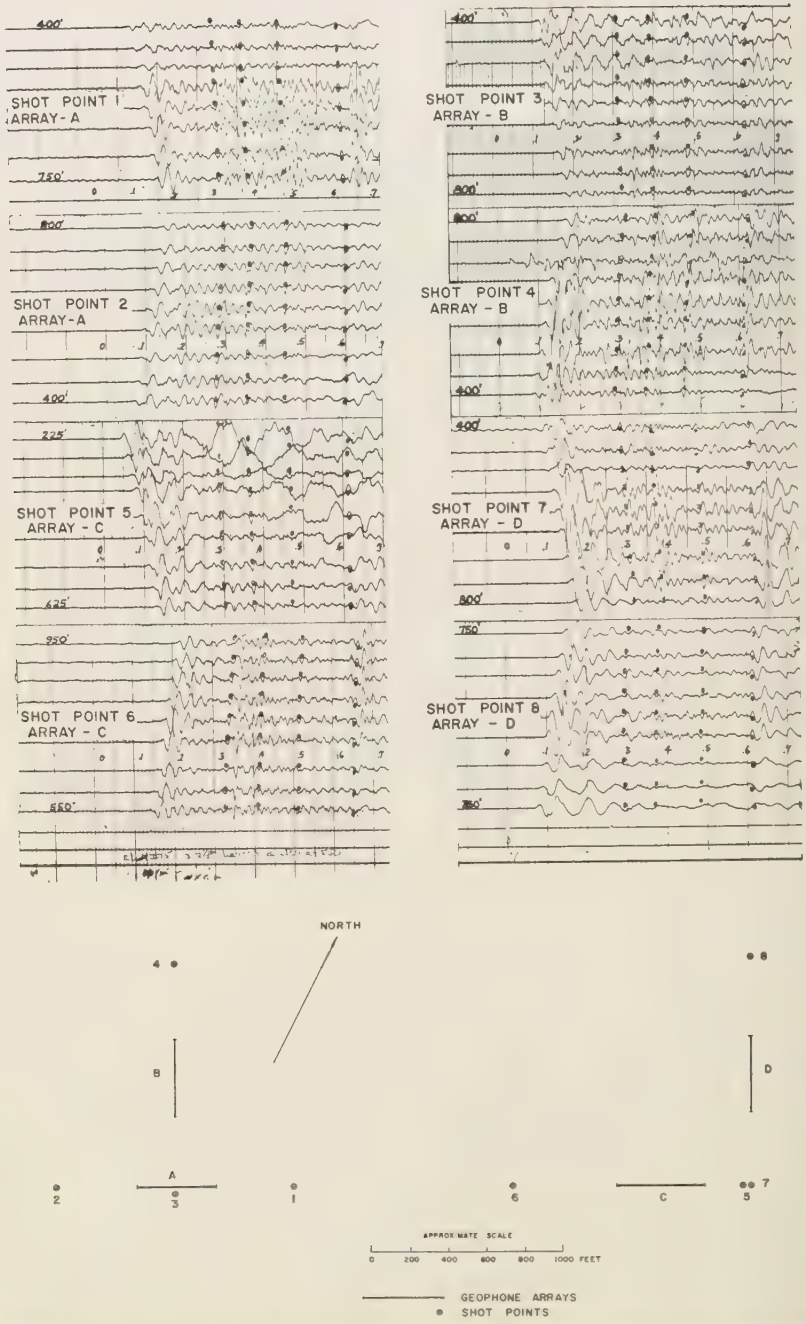


FIG. 6.—Beach reflection records and shot plan. Full circles on records indicate reflections that can be correlated on all recordings shown

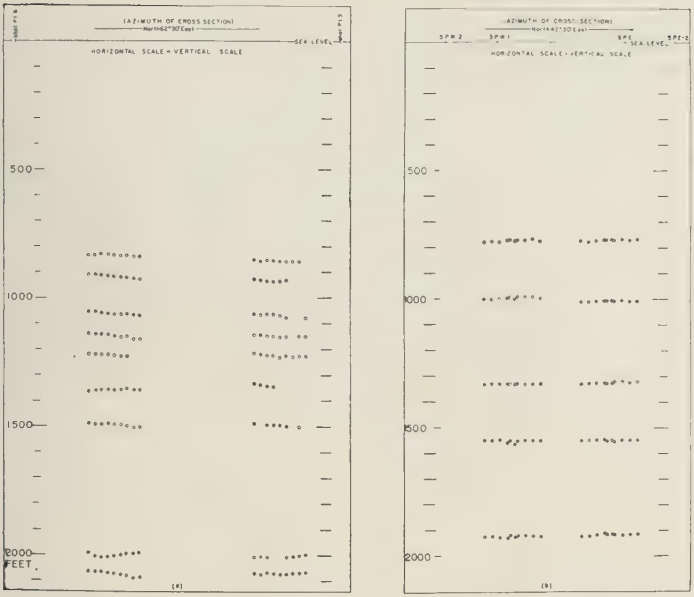


FIG. 7—Point-plotted reflection depth profiles; lagoon profile (a), beach profile (b)

By plotting the quantity $(T_A^2 - T_B^2)$ for each detector between shotpoints A and B as a function of x, we can define a straight line. The slope of this line may be used to compute $\bar{V}$. Better approximations to the interval velocities may then be computed from $\bar{V}$ by use of equation (12) on p. 73 of *Dix* [1955]. Errors introduced by neglecting the dip effect are very small, since the dip, θ, enters the slope term as $\sin^2 \theta$.

The method just described reduced errors due to refraction of rays in multilayered media. The dip may be calculated by successive approximations, but this was unnecessary for the data reported here. The results were classified according to reliability because there were indications of multiple reflection effects and minor effects due to propagation irregularities [*Dix*, 1952, Chapter 11]. The most reliable data were combined and are presented as vertical velocity distributions in Figure 8, along with the velocity distributions obtained from the refraction profiles and the borehole measurements.

Velocity distributions

In interpreting the seismic reflection and refraction data reported in this paper, many simplifying assumptions were necessary. Of these assumptions, the introduction of a simple model consisting of relatively few parallel homogeneous seismic layers for the range of wave lengths, seismic velocities, and rock densities involved is probably the least justified. As a direct result of assuming this, it is necessary to consider the effect of anisotropy when comparing the various velocity data of Figure 7. The importance of this effect has been clearly shown by *Hagedoorn* [1954]. By following his reasoning we conclude that the vertical velocities for the borehole and for the reflections (Fig. 8) are mean values, and also that the refraction profiles show horizontal velocities which are maximum values.

The borehole interval velocities show four significant velocity inversions (Fig. 2). Two zones of particularly low speed (4600 ft/sec and 4320 ft/sec) are believed to be due to inclusions of large amounts of organic matter (lignite) indicated in the driller's log. Because the low speed layers are 'masked' in refraction profiling, one would expect the refraction boundaries plotted in Figures 3 and 5 to be deeper than the borehole indicates. However, Figure 8 shows the depth to be about the same. This may be due to the presence of high speed layers which are also 'masked.' This effect may cancel or even reverse

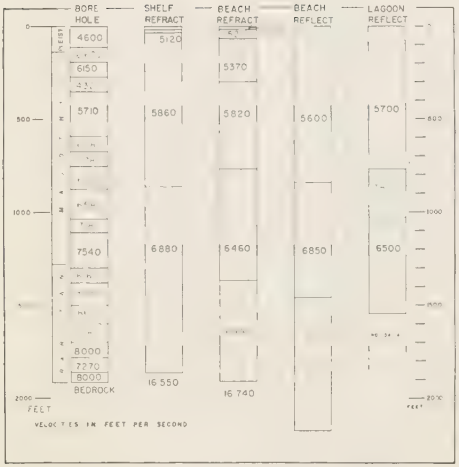


FIG. 8—Summary of seismic velocity distributions

the effect of the 'masked' low speed layers in the depth calculations.

The reflection data have the inherent disadvantage that basement velocity is not represented. In fact basement reflections are difficult to obtain on the continental shelf [*Ewing* and others, 1937]. The geological significance of the reflection from within basement, as shown in the beach reflection profile of Figure 6, is not clear. Furthermore, the strong development of a late low frequency event several hundredths of a second after basement arrival time, according to the borehole log, with small reflection 'step-out' is puzzling. Analysis of the records has not shown this to be a multiple reflection. Another difficulty with calculation of velocities from reflection data is the need for many accurate reflection-time measurements. *Swan* and *Becker* [1952] and *Widess* [1952] have also considered this problem.

Comparison of seismic profile velocities with interval velocities in the borehole for all of the layers shown in Figure 7 gives differences of one to four per cent, with the exception of one anomalous case. The indication of a 250-ft thick layer with a velocity of 7800 ft/sec in the lagoon profile is believed to be inaccurate because of the difficulty in picking beginnings of the reflections from the top and bottom of the layer. The error in this case is twenty-four per cent. If we do not include this exceptionally large error, our seismic profile velocities may be accepted as remarkably good approximations to directly measured subsurface velocities.

Acknowledgments—This work was carried out by the personnel of Hudson Laboratories of Columbia University under Contract N6-ONR-27135 with the Office of Naval Research. It is a pleasure to acknowledge the cooperation of R. A. Frosch, Director of Hudson Laboratories, I. Tolstoy, M. V. Brown, and Roberto Frassetto in this project. Special acknowledgment is due to J. E. Upson of the U. S. Geological Survey, Mineola, N. Y., for his financial help and cooperation in the well logging program, and to N. M. Perlmutter, also of the U.S.G.S., for observing the drilling operations and analyzing the cuttings.

REFERENCES

BECKMAN, W. C., A. C. ROBERTS, AND B. LUSKIN, Sub-bottom depth recorder survey of western Long Island Sound, *Trans. Am. Geophys. Union*, **39**, 505–506, 1958.

BROWN, M. V., J. NORTHROP, R. FRASSETTO, AND L. GRABNER, Seismic-refraction profiles on the continental shelf, in preparation.

CARLSON, R. O., AND M. V. BROWN, Seismic-refraction profiles in the submerged Atlantic Coastal Plain near Ambrose Lightship, *Bull. Geol. Soc. Am.* **66**, 969–976, 1955.

DIX, C. H., *Seismic prospecting for oil*, Harper and Bros., New York, 1952.

DIX, C. H., Seismic velocities from surface measurements, *Geophysics*, **20**, 68–86, 1955.

DOMZALSKI, W., Some problems of shallow reflection investigations, *Geophys. Prosp.* **4**, 140–166, 1956.

EWING, M., A. P. CRARY, AND H. M. RUTHERFORD, Geophysical investigations in the emerged and submerged Atlantic Coastal Plain, Part I: Methods and Results, *Bull. Geol. Soc. Am.*, **48**, 753–802, 1937.

EWING, M., J. L. WORZEL, N. C. STEENLAND, AND F. PRESS, Geophysical investigations in the emerged and submerged Atlantic Coastal Plain, Part V: Woods Hole, New York, and Cape May sections, *Bull. Geol. Soc. Am.*, **61**, 877–892, 1950.

HAGEDOORN, J. G., A practical example of an anisotropic velocity layer, *Geophys. Prosp.*, **2**, 52–60, 1954.

MCMILLEN, J. H., The velocity of dilatation and Rayleigh waves in metal bars, *J. Acoust. Soc. Am.* **18**, 190–199, 1946.

OLIVER, J. E., AND C. L. DRAKE, Geophysical investigations in the emerged and submerged Atlantic Coastal Plain, Part VI: The Long Island Area, *Bull. Geol. Soc. Am.*, **62**, 1287–1296, 1951.

STRICK, E., W. L. ROEVER, AND T. F. VINING, Theoretical and experimental investigation of a pseudo-Rayleigh wave, (abstract), Proc. Sound I.C.A. Congress, *Sound and Man*, p. 186, 1956.

SUTER, R., W. DELAGUNA, AND N. M. PERLMUTTER, Mapping of geologic formations and aquifers of

Long Island, N. Y., Dept. Conserv., *Water Power and Control Comm. Bull. GW-18*, 212 pp., 1949.

Swan, B. G., and A. Becker, Comparison of velocities obtained by deltatime analysis and well velocity surveys, *Geophysics*, **17**, 575, 1952.

Tolstoy, I., Resonant frequencies and high modes in layered wave guides, *J. Acoust. Soc. Am.*, **28**, 1182–1192, 1956.

Tolstoy, I., Shallow water test of the theory of layered wave guides, *J. Acoust. Soc. Am.*, **30**, 348–361, 1958.

Widess, M. B., Salt solution, a seismic velocity problem in western Anadarko Basin, Kansas, Oklahoma, Texas, *Geophysics*, **17**, 481–504, 1952.

(Manuscript received September 24, 1956; presented at the Thirty-Ninth Annual Meeting, Washington, D. C., May 5, 1958.)

A General Theory of the Unit Hydrograph

James C. I. Dooge

Civil Engineering Department
University College
Cork, Ireland

Abstract—By the single assumption that the reservoir action in a catchment can be separated from translation, the general equation of the unit hydrograph is shown to be

$$u(0, t) = \frac{V_0}{A} \int_0^{A(t)} \frac{\delta(t - \tau)}{\Pi(1 + K_i D)} \cdot i \cdot dA$$

This is simplified by two further simple assumptions to give

$$u(0, t) = \frac{V_0}{T} \int_0^{t \leq T} P(m, n - 1) \cdot \omega(\tau') \cdot dm$$

which can be conveniently calculated.

List of special symbols

A—area of catchment
cross section of channel (Section 3)
$a(x)$—translation time
$I(t)$—inflow rate
K—delay time of reservoir storage
M—*-th* moment of the unit hydrograph
N—total number of reservoirs in the catchment
m—dimensionless time variable
$n(\tau)$—distribution of reservoirs in the catchment
$P(m, n)$—Poisson probability function
$Q(t)$—outflow rate
$S(t)$—storage volume
T—maximum translation time in catchment
$U(0, t)$—instantaneous unit hydrograph
$U(D, t)$—finite period unit hydrograph
V_0—volume of rainfall excess (outflow)
$\delta(0)$—Dirac-delta function
Π—product of similar terms to be taken
τ—total translation delay time between point and outlet
$\omega(\tau)$—ordinate of time-area-concentration curve
i—$\frac{\text{local rainfall intensity}}{\text{average rainfall intensity}}$

Introduction

Since its first proposal over 25 years ago [*Sherman*, 1932], the unit hydrograph approach to stream flow prediction has developed into one of the most powerful tools of applied hydrology. It has, however, retained an empirical character, and no general theoretical basis for the method has been evolved. The absence of a general theory of the unit hydrograph has limited the scope of the method and made it dependent on personal judgment in its practical application.

The present paper proposes a general theory which should help to remove many of the subjective elements from unit hydrograph analysis, and also to release the problem of synthesis from its present dependence on empirical relationships derived from localized data. To be worthwhile, a theory should enable us to predict the effect on the unit hydrograph of such variables as the shape of the catchment, the speed of travel of flood waves in the catchment, the amount and distribution of both channel storage and overbank storage, unequal rainfall distribution over the catchment, the size and slope of the catchment, and so on. It would be gratifying, also, if such a theory could (a) provide a simple objective method of deriving a unit hydrograph from complex storms; and (b) indicate a small number of physically significant parameters which could be correlated with catchment characteristics to form the basis of a universally applicable system of unit hydrograph synthesis.

Zoch [1934, 1936, 1937] made a notable attempt to formulate a general physical theory of stream flow based on the assumption that at any time the rate of discharge is proportional to the amount of rainfall remaining with the soil at that time. Using this assumption, he analysed the runoff

due to rainfall of finite duration and uniform rate, obtaining equations for the four separate segments of the hydrograph. Zoch solved these equations for two simple cases, where the time-area-concentration curve is rectangular or triangular. He indicated as the main difficulty in the solution of other cases the integration of $\omega(x)e^{Kx}$ where $\omega(x)$ is the ordinate of the time-area-concentration curve and K is a constant; he suggested the use of series approximation or numerical integration in such cases.

Clark [1945] suggested that the unit hydrograph for instantaneous rainfall could be derived by routing the time-area-concentration curve through a single element of linear reservoir storage. Physically this is equivalent to Zoch's formulation, the equations being simplified by reducing rainfall duration to zero, and the reservoir routing being a form of numerical integration of $\omega(x)e^{Kx}$. The hydrology group of the Irish Office of Public Works found that in many practical cases the smoothing involved in routing was sufficient to permit the replacement of the tediously derived time-area-concentration curve by an isosceles triangle [*O'Kelly*, 1955]. The present author in a thesis study discussed the physical basis of the Clark approach and the O'Kelly simplification [*Dooge*, 1956].

A further hypothesis was put forward by *Nash* [1957], more as the basis of a convenient two-parameter synthetic method than as a general theory. He suggested that the instantaneous unit hydrograph could be derived by routing the instantaneous rainfall through a series of successive linear reservoirs of equal delay time.

In the present paper, a general equation for the unit hydrograph is derived from the single physical assumption that the reservoir action which takes place in the catchment can be separated from the translatory action and lumped in a number of reservoirs unrestricted in number, size, or distribution. This general equation can be converted from a surface integral to a simple integral by assuming that above each confluence in the catchment the reservoirs are equally distributed for equal lengths of tributary. The complexity of the computation required is greatly reduced if the idealised reservoirs in the catchment are assumed equal, since in this case the ordinates of the unit hydrograph can be obtained by integrating the product of an ordinate of the time-area-concentration curve and an ordinate of the Poisso probability function (which has been tabulated Even with this assumption of equal reservoir the result is extremely general and flexibl containing each of the previous theoretic approaches as a very special case.

Unit hydrograph principles

All unit hydrograph procedures (analytical synthetic) are based on two fundamenta principles:

1. Invariance—the hydrograph of surfac runoff from a catchment due to a given patter of rainfall excess (that is, rainfall minus infiltr tion and similar losses) is invariable.

2. Superposition—the hydrograph resultin from a given pattern of rainfall excess can b built up by superimposing the unit hydrograph due to the separate amounts of rainfall exce occurring in each unit period. This includes th principle of proportionality, by which the ord nates of the hydrograph are proportional to th volume of rainfall excess.

Further assumptions are often made f practical convenience, but they are not essentia Thus the starting point for Sherman's origin work on the subject was the assumption that a floods due to rainstorms of a given duration ra off in the same amount of time. This use of finite time base for the unit hydrograph is n essential and is only physically reasonable f cases where the storage is distributed even throughout the catchment.

For the purpose of the present study th following definitions are used. The instantaneou unit hydrograph for a catchment, $u(0, t)$ is th hydrograph of surface runoff at the outlet fro the catchment due to a finite volume of rainfa excess, V_0, falling in an infinitesimally short tim Such a rainfall excess is written in terms of th δ-function as $V_0 \cdot \delta(0)$. A D-period unit hydr graph, $u(D, t)$, is the hydrograph of surfac runoff at the outlet due to a rainfall excess volume, V_0, distributed uniformly throughout period, D. In both cases the areal distribution rainfall excess throughout the catchment assumed to follow a constant pattern from stor to storm, but this pattern is not necessarily one uniform distribution. The conversion from a instantaneous unit hydrograph to one of finit duration can be easily made, since by supe position

$$u(D, t) = \frac{1}{D}\int_{t-D}^{t} u(0, t)\cdot dt$$

Unless the contrary is stated, all unit hydrographs discussed in the present paper are instantaneous unit hydrographs whose ordinates have the dimension of discharge rate per unit area.

The presence of the principle of superposition implies that any theory of the unit hydrograph must be a linear one. The processes involved in the conversion of rainfall excess to surface runoff at the outlet must all be linear if the unit hydrograph theory is to hold exactly, since the presence of a single non-linear element would be sufficient to destroy the principle of superposition. It is clear that any general theory must be a linear one, and that cases of catchments containing non-linear elements must be dealt with by some process of linearization.

The key problem, therefore, is to determine the equation of the instantaneous unit hydrograph for a catchment containing only linear elements. The corresponding problem in applied mathematics is the determination for a linear system of the Green's function, that is, the response of the system to the δ-function or unit impulse. The first step towards the solution of the hydrological problem is to examine the response of simple linear elements to an instantaneous inflow.

Once the instantaneous unit hydrograph is known, the runoff due to any given rainfall pattern can be found by convolution,

$$q = i(t) * u(0, t)$$

$$= \int_0^t u(0, t - \tau)\cdot i(\tau)\cdot d\tau$$

In practice this final calculation is most conveniently made by using a distribution graph, or by using a movable strip technique.

Linear channels and reservoirs

The process of converting rainfall excess into surface runoff is a mixture of translation and reservoir action. The first step to be taken is to examine the cases of pure reservoir action and pure translation in a linear catchment.

A linear reservoir is one in which the storage is directly proportional to the outflow:

$$S(t) = KQ(t)$$

the constant K having the dimension of time and being equal to the average delay time imposed on an inflow by the reservoir. If the above relationship is combined with the storage equation

$$I(t) - Q(t) = \frac{d}{dt} S(t)$$

we get

$$I(t) - Q(t) = K\frac{d}{dt} Q(t)$$

which can be written in operational form as

$$(1 + KD)Q(t) = I(t)$$

where D is the differential operator. This equation has the solution

$$Q(t)e^{t/K} = \frac{1}{K}\int I(t)e^{t/K}\, dt + \text{constant}.$$

For any given inflow this equation can be solved either analytically or numerically. For an instantaneous inflow the outflow from a linear reservoir is given by

$$Q(t) = e^{-t/K}\left[\int \frac{V_0}{K} e^{t/K}\, \delta(0)\cdot dt + \text{const.}\right]$$

$$= \frac{V_0}{K} e^{-t/K}, \qquad t > 0$$

so that the response of a linear reservoir to the δ-function is a sudden jump at the instant of inflow to a finite outflow followed by an exponential decline approaching infinity.

In order to discuss the problem of translation, it is convenient to introduce the concept of a linear channel. Such a channel is analogous to the idea of a linear reservoir whose storage outflow curve is a straight line. A channel unaffected by backwater has a definite rating curve at every point of the channel. A linear channel is defined as a reach in which the rating curve at every point is a linear relationship between discharge and area. This implies that at any point the velocity is constant for all discharges, but may vary from point to point along the reach. If the linear rating is written as

$$A = a'(x)\cdot Q$$

this equation can be applied to any point along the linear channel. This relationship can be

combined with the continuity equation

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = 0$$

giving the equation

$$\frac{\partial Q}{\partial x} + a'(x) \cdot \frac{\partial Q}{\partial t} = 0$$

which has the solution

$$Q[t - a(x)] = \text{constant}.$$

This solution corresponds to the case of pure translation. It indicates that a linear channel will translate any inflow hydrograph without change of shape. This is in contrast with the non-linear case in which only one inflow hydrograph is capable of uniform translation, that is, the special case of uniformly progressive flow. For the case of instantaneous inflow at the upstream end of a linear channel, the flow at any other point of the channel is given by

$$Q = V_0 \cdot \delta(a)$$

Combinations of channels and reservoirs

If a linear reservoir and a linear channel are placed in series, the order in which they occur is immaterial, since the translation due to the linear channel merely involves the shifting of a time scale. The outflow from two such elements in series due to an instantaneous inflow $V_0 \cdot \delta(0)$ is given by

$$t \geq a \qquad Q = \frac{V_0}{K} e^{-(t-a)/K}$$

where K is the delay time due to the linear reservoir and a is the translation time due to the linear channel.

If a number of linear channels are placed in series, the effect is merely one of translation by an amount equal to the sum of the translation times of the individual linear channels.

If a number of unequal linear reservoirs occur in series, the outflow from one is the inflow to the next, so that the equation is

$$(1 + K_1 D)(1 + K_2 D) \cdots (1 + K_n D) Q_n(t) = I(t)$$

or

$$Q_n(t) = \frac{1}{(1+K_1 D)(1+K_2 D)\cdots(1+K_n D)} I(t)$$

which can be written as

$$Q_n(t) = \frac{1}{\Pi_1^n (1 + K_i D)} \cdot I(t)$$

For the instantaneous inflow this reduces to

$$t > 0 \qquad Q_n(t) = \frac{1}{\Pi(1 + K_i D)} \cdot \{0\}$$

which is easily solved, since the inverse operator is already divided into factors.

If all the linear reservoirs are unequal, the outflow must be of the form

$$Q_n(t) = C_1 e^{-t/K_1} + C_2 e^{-t/K_2} + \cdots + C_n e^{-t/K_n}$$

Since the flow from any one of the reservoirs decreases to zero after an infinite time, we can write for each reservoir

$$\int_0^\infty Q_i \cdot dt = V_0 \qquad i = 1, 2 \cdots n$$

These n equations form the boundary conditions which enable us to evaluate the n unknown coefficients $C_1 \cdots C_n$. The insertion of these boundary conditions gives the set of equations

$$\int_0^\infty Q_n \, dt = V_0$$

i.e.

$$\int_0^\infty (C_1 e^{-t/K_1} + C_2 e^{-t/K_2} + \cdots C_n e^{-t/K_n}) \cdot dt = V_0$$

$$\int_0^\infty Q_{n-1} \cdot dt = V_0$$

$$\int_0^\infty (1 + K_n D) Q_n \cdot dt = V_0$$

$$\int_0^\infty Q_n \cdot dt + \int_0^\infty K_n \frac{d}{dt} Q_n \cdot dt = V_0$$

$$V_0 + K_n [Q_n]_0^\infty = V_0$$

i.e.

$$[C_1 e^{-t/K_1} + C_2 e^{-t/K_2} + \cdots C_n e^{-t/K_n}]_0^\infty = 0$$

etc. etc.

which reduce to the algebraic equations

$$C_1K_1 + C_2K_2 + \cdots + C_nK_n = V_0$$
$$C_1 + C_2 + \cdots + C_n = 0$$
$$\frac{C_1}{K_1} + \frac{C_2}{K_2} + \cdots + \frac{C_n}{K_n} = 0$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots = 0$$
$$\frac{C_1}{K_1^{n-2}} + \frac{C_2}{K_2^{n-2}} + \cdots + \frac{C_n}{K_n^{n-2}} = 0$$

On solving these, the general equation can be written as

$$\frac{Q_n(t)}{V_0} = \frac{K_1^{n-2}e^{-t/K_1}}{\Pi(K_1 - K_i)} + \frac{K_2^{n-2}e^{-t/K_2}}{\Pi(K_2 - K_i)} + \cdots + \frac{K_n^{n-2}\cdot e^{-t/K_n}}{\Pi(K_n - K_i)}$$

so that the outflow from a chain of unequal linear reservoirs can be expressed as a sum of n terms, each of which is a simple exponential decay curve. This does not furnish us with an easily manipulated tool for analysis and subsequent synthesis, since the separation of exponentials by numerical analysis is an unsatisfactory process [*Lanczos*, 1957]. If some of the n linear reservoirs are equal, repeated roots are obtained. This complicates some of the individual terms without reducing the number of terms in the series.

In the special case where all of the n linear reservoirs are equal, a remarkable simplification is obtained. In this case we have

$$Q_n(t) = (1 + KD)^{-n}\{0\}$$
$$= (C_1 + C_2t + C_3t^2 + \cdots + C_nt^{n-1})e^{-t/K}$$

Insertion of the boundary conditions eventually results in the following

$$C_1 = C_2 = C_3 = \cdots = C_{n-1} = 0$$
$$C_n = \frac{V_0}{K^n(n-1)!}$$

so that the outflow is given by

$$Q_n(t) = \frac{V_0}{K}\frac{(t/K)^{n-1}e^{-t/K}}{(n-1)!}$$

which shows that the outflow is represented by a single term. The fact that the term is complicated is of no practical significance, since it is identical with the Poisson probability function, values of which have been tabulated by *Pearson* [1930], and *Molina* [1942]. This fact is taken advantage of later in the paper to facilitate a numerical solution of the general equation of the unit hydrograph. The outflow from n equal reservoirs can, therefore, be written conveniently as

$$Q_n(t) = \frac{V_0}{K}P\left(\frac{t}{K}, n-1\right)$$

where P is the Poisson probability function.

One final case is of special interest. If a total delay time of K is assumed, the effect on the outflow hydrograph of varying the value of n gives some interesting results. The maximum outflow occurs at a time $(n-1)K$ and has the value

$$Q_{max} = \frac{V_0}{K}\cdot P(n-1, n-1)$$

As n increases, the peak increases and other values decrease until for n equal to infinity the outflow is $V_0\cdot\delta(n-1\cdot K)$. This result indicates that if a finite total delay time is divided up among a large number of equal linear reservoirs, the effect on an instantaneous inflow is equivalent to translation by an amount equal to the total delay time.

Thus we can look on (1) reservoir action as being related to concentrated storage and (2) translatory action as related to completely distributed storage.

A GENERAL EQUATION FOR THE UNIT HYDROGRAPH

The linear components already discussed can be combined to represent an ideal linear catchment and an equation written for the unit hydrograph for such a catchment. This is done by assuming that any linear catchment can be represented by an ideal linear catchment in which the storage (both channel and overbank) is either concentrated or completely distributed; that is, the ideal catchment is drained by a network of channels composed of a complex system of linear channels and linear reservoirs placed in series. This lumping of the reservoir action is the only assumption made in deriving the general equation (apart from the necessary assumption of linearity without which no true unit hydrograph would exist).

The first step is to consider the outflow from

an element of area in the idealized catchment. Since the whole system is linear, the contribution of any portion can be considered in isolation. To reach the outlet the rainfall excess falling on any element of area must pass through a linear reservoir representing either overland flow or interflow, and then pass through the chain of linear channels and linear reservoirs representing the drainage system between the point under consideration and the outlet. Since the order of operations is immaterial, the flow may be considered as first passing through a succession of linear channels and then through the series of linear reservoirs appropriate to the portion of the catchment under examination. For a rainfall excess $I(t)$ on an element of catchment ΔA the resulting outflow is given by

$$\Delta Q(t) = \frac{1}{\Pi_1^n(1 + K_i D)} \cdot \{I(t - \tau)\} \cdot \Delta A$$

where

τ = total translation time between the element and the outlet

and

$K_1 \cdots K_n$ = storage delay times of reservoirs between the element and the outlet.

From this it follows that the contribution of the element of area ΔA to the instantaneous unit hydrograph is given by

$$t < \tau \qquad \Delta u(0, t) = 0$$

$$t \geq \tau \qquad \Delta u(0, t) = \frac{1}{\Pi(1 + K_i D)} \; V_0 \delta(t - \tau) \; \frac{\Delta A}{A}$$

If the value of τ and the values of the n storage delay times $K_1 \cdots K_n$ are known, this equation can be calculated as indicated in the preceding section.

The total runoff from the catchment after a time t will clearly be the sum of the contributions of the individual elements of area. It is obvious that those areas, where the total translation time τ is greater than the elapsed time t since the start of rainfall excess, will contribute nothing to the runoff at the outlet, as the initial effect of the rainfall excess will still be in transit. Summation over the area that is contributing gives us the following expression for the instantaneous unit hydrograph:

$$u(0, t) = \sum_{A(0)}^{A(t)} \frac{1}{\Pi(1 + K_i D)} \cdot V_0 \delta(t - \tau) \cdot i \cdot \frac{\Delta A}{A}$$

If the integral is taken in the Stieltjes sense to allow for discontinuities, we can write this as

$$u(0, t) = \frac{V_0}{A} \int_0^{A(t)} \frac{\delta(t - \tau)}{\Pi(1 + K_i D)} . i . dA$$

in which

$u(0, t)$ = ordinate of the instantaneous unit hydrograph
V_0 = volume of runoff
A = area of catchment
t = time elapsed
τ = translation time
$K_1 \cdots K_n$ = storage delay times
$i(A)$ = ratio of local rainfall intensity to average over catchment.

The above equation is the general equation for the unit hydrograph of an ideal linear catchment in which translatory action and reservoir action are separated. If the rainfall intensity is constant over the catchment, then $i = 1$ and can be omitted. It is reasonable to suppose that any catchment suitable for unit hydrograph analysis can be represented by an equivalent ideal linear catchment consisting of linear channels and linear reservoirs so arranged as to give the same unit hydrograph as the natural catchment to the required degree of accuracy. Accordingly, the above equation is proposed as the general equation of the instantaneous unit hydrograph.

Methods of solution

The exact solution of the general equation derived in the last section would be a tedious business. It would require the evaluation for each element of the catchment area of the n-term series (generated by the n reservoirs lying between the element and the outlet) for all the values of $(t - \tau)$, and then the combination of the contributions from each element to obtain the complete unit hydrograph. From a practical viewpoint, the accuracy of the basic data renders the tedium of such a computation pointless, and in any case the accuracy required in applied hydrology makes this tedium unnecessary. Since the instantaneous unit hydrograph, which is itself an integral, is further integrated to obtain the actual hydrograph, small differences in catchment characteristics will not be reflected in the hydro-

graph of surface runoff due to the double smoothing involved. This suggests that the general equation of the unit hydrograph can be made tractable by the use of carefully chosen physical assumptions without undue loss of accuracy.

It is obviously desirable to transform the general equation from a surface integral to a single integral. This can be done as follows. For the purpose of solving the general equation, any point in the catchment is characterized by its translation time τ from the outlet and by the set of reservoirs with delay times $K_1 \cdots K_n$ lying between the point in question and the outlet. Imagine a 'contour' joining all points in the catchment which are separated from the outlet by the same translation time; such a line is termed an isochrone. Such isochrones cannot cross one another, cannot close, and can only originate or terminate on the boundary of the catchment. In general, it is possible for the points on an individual isochrone to have different characteristic 'reservoir chains' lying between the point and the outlet. This is illustrated for a simple case in Figure 1. Natural

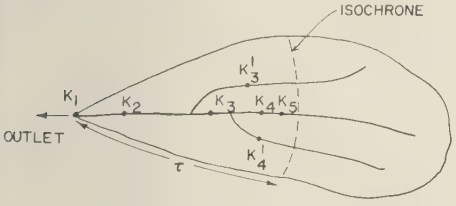


FIG. 1—General distribution of reservoirs

catchments, however, seem to conform to certain laws of equilibrium [*Horton*, 1945]. Thus the assumption that all points on the same isochrone have the same reservoir chain would not be unduly restrictive. This assumption is shown in Figure 2; it is clear that all the tributaries can be folded onto the main river to give a single chain of reservoirs as shown in Figure 3, the inflow at any point being proportional to the length of the isochrone ($dA/d\tau$) cutting the main river at that point. In order to take advantage of this assumption for non-uniform rainfall distribution, it is necessary to average the rainfall intensity along each isochrone, so that the intensity varies only with distance from the outlet.

The physical assumption made in the last

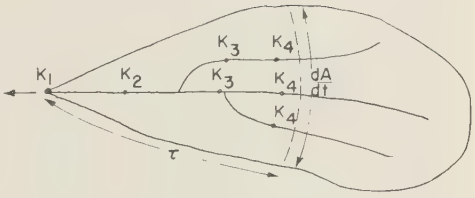


FIG. 2—Uniform distribution of reservoirs

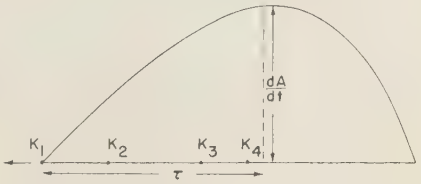


FIG. 3—Folding of tributaries onto main river

paragraph modifies the general equation

$$u(0,\ t) = \frac{V_0}{A} \int_0^{A(t)} \frac{\delta(t - \tau)}{\Pi(1 + K_i D)} . i . dA$$

to the following

$$u(0,\ t) = \frac{V_0}{A} \int_0^{t \leq T} \frac{\delta(t - \tau)}{\Pi(1 + K_i D)}\ i\ \frac{dA}{d\tau}\ d\tau$$

where

$dA/d\tau$ = the length of the isochrone (i.e., the ordinate of the time-area curve)

T = the classical "time of concentration" (i.e., the maximum translation time in the catchment).

This can be written

$$u(0,\ t) = \frac{V_0}{T} \int_0^{t \leq T} \frac{\delta(t - \tau)}{\Pi(1 + K_i D)}\ \omega\left(\frac{t}{T}\right)\ d\tau$$

where

$\omega\left(\frac{\tau}{T}\right)$ = ordinate of the dimensionless time-area concentration curve, adjusted for variation in rainfall intensity

or in compact dimensionless form

$$\frac{uT}{V_0} = \int_0^{t' \leq 1} \frac{\delta(t' - \tau')}{\Pi(1 + K_i D)}\ \omega(\tau')\ d\tau'$$

which is a single integral since the chain of reservoirs is defined uniquely for each value of τ. It should be kept in mind that translation may

include distributed storage in the form of continuous overbank flow. Hence in the present theory the translation time is based on the velocity of flow from the whole cross section and does not exclude overbank flow from the calculation. This is in contrast with the classical rational method, where time of concentration is based on bankfull flow conditions.

The integral given above can be further simplified by assumptions in relation to the time-area-concentration curve. A complex ω-curve can be broken down into simpler elements for which solutions are easily found, and then these solutions can be recombined to give the complex solution for the complete curve. In particular, if the curve is replaced by a series of straight lines each of which can be expressed as

$$\omega(\tau') = C_1 + C_2\tau' \qquad a < \tau < b$$

then the unit hydrograph is given by

$$\frac{uT}{V_0} = C_1 \int_a^b \frac{\delta(t-\tau)}{\Pi(1+K_iD)}\,d\tau + C_2 \int_a^b \frac{\delta(t-\tau)}{\Pi(1+K_iD)}\,\frac{\tau}{T}\,d\tau$$

In this way the unit hydrograph for any polygonal $\omega(\tau')$ can be built up from the basic unit hydrographs for $\omega = 1$ and $\omega = \tau'$. In many cases it will be simpler to solve the complex case by numerical integrations rather than by combining the basic solutions over short segments.

In order to compute an actual unit hydrograph from the general equation, the size and distribution of the linear reservoirs must be known or assumed. In general, these reservoirs may be of any size and placed at any given translation time from the outlet. The Zoch and Clark methods assume, in effect, that there is a single reservoir in the catchment. Using this assumption, the unit hydrograph can be computed in less than half an hour, even for the case of a complex time-area-concentration curve. If more reservoirs are inserted in the idealized catchment, the computational work is increased and would become practically impossible for a large number of unequal reservoirs. If, however, the reservoirs are all taken as equal, there is a great simplification and the number of reservoirs can be increased without extra work up to the limit of the tables used. In the case of a number of equal reservoirs, the computation takes two to three hours.

The solutions for these various assumptions regarding the linear reservoirs are discussed in greater detail in the remaining sections of the paper.

Existing solutions

The various theoretical solutions so far proposed for relating rainfall to runoff on a linear basis are easily derived as special cases of the general equation

$$\frac{uT}{V_0} = \int_0^t \frac{\delta(t-\tau)}{\Pi(1+K_iD)} \cdot \omega(\tau') \cdot d\tau$$

The insertion of particular values in the above equation gives the solutions proposed in the Rational Method (1851), *Zoch* [1934], *Clark* [1945], and *Nash* [1957].

The Rational Method first proposed by Mulvany in 1851 [*Dooge*, 1957] treats the problem of runoff as one of translation only. This assumption corresponds to

$$K_1 = K_2 = \cdots = K_n = 0$$

which gives

$$\frac{uT}{V_0} = \int_0^t \delta(t-\tau)\omega(\tau')\,d\tau = \omega(t')$$

so that the instantaneous unit hydrograph has the same shape as the time-area-concentration curve. In the original form of the Rational Method, the critical storm duration was taken as equal to T, the time of concentration of the catchment. In the modified rational method, the critical storm is taken so as to maximize the product of the rainfall intensity and the peak of the finite period unit hydrograph.

Zoch [1936] and *Clark* [1945] both assumed that the runoff was routed through a single linear reservoir. Zoch assumed the storage to occur in the soil, while Clark more correctly regarded it as channel storage. This assumption corresponds to

$$K_1 = K$$

$$K_2 = K_3 = \cdots = K_n = 0$$

which when inserted in the general equation gives

$$\frac{uT}{V_0} = \int_0^t \frac{\delta(t-\tau)}{1+KD}\,\omega(\tau')\,d\tau$$

$$= \int_0^t \frac{1}{K}\,e^{-(t-\tau)/K}\,\omega(\tau')\,d\tau$$

$$= \frac{e^{-t/K}}{K}\int_0^t e^{\tau/K}\,\omega(\tau')\,d\tau$$

The latter equation is identical with the solution obtained when a time-area-concentration curve $\omega(\tau')$ is routed through a reservoir of delay time K in accordance with Clark's method. In practice, such a routing can be carried out very rapidly by using the coefficient form of the Muskingum method to obtain the instantaneous unit hydrograph. The finite-period unit hydrograph can be readily obtained by taking means and entering these in a further column; if required, a distribution graph can be obtained by using a further column [*Dooge*, 1956].

The instantaneous unit hydrographs computed on the assumption of a single reservoir for catchments with rectangular and triangular $\omega(\tau)$ curves are shown in Figures 4 and 5. The range of K/T values used corresponds with that normally found in natural catchments. It is found that the variation of the peak of the unit

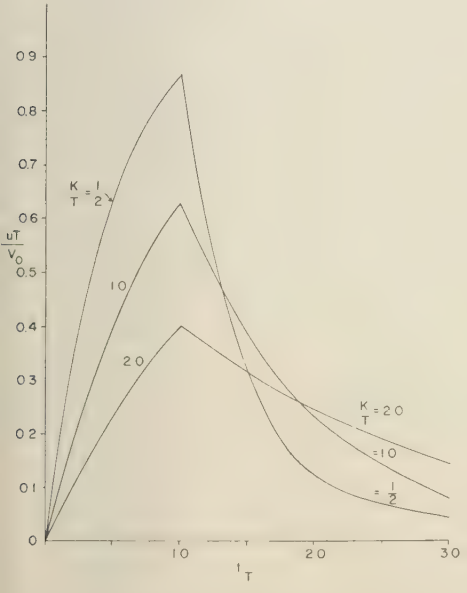


FIG. 4—Hydrograph for rectangular catchment

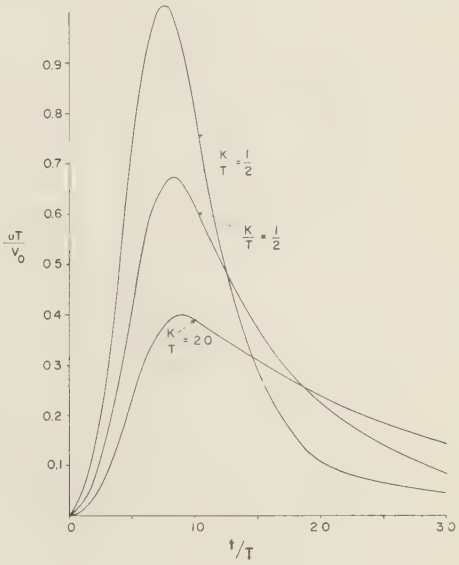


FIG. 5—Hydrograph for triangular catchment

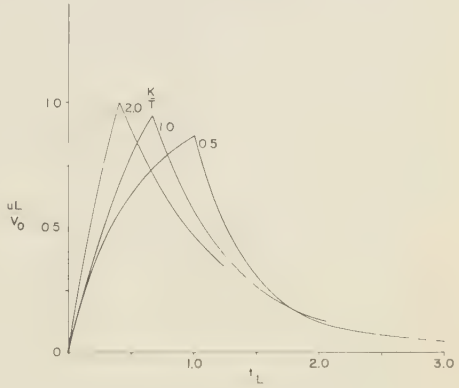


FIG. 6—Hydrograph in terms of lag (rectangular)

hydrograph is greatly reduced if the plotting is in terms of the lag (or first moment) L of the unit hydrograph. This plotting which is for uL/V_0 versus t/L is shown for rectangular and triangular catchments respectively in Figures 6 and 7. The remarkable constancy of the dimensionless peaks is shown in Table 1. It is clear from the above table that instantaneous unit hydrographs derived by the O.P.W. Method [*O'Kelly*, 1955], that is, by routing an isosceles triangle through a linear reservoir, will have the property that the product of the peak and

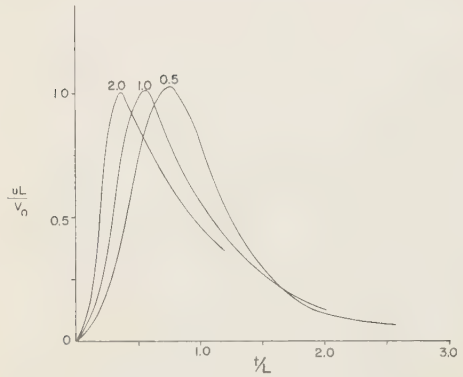


FIG. 7—Hydrograph in terms of lag (triangular)

the lag equal the volume within a few per cent.

Nash [1957] used a different approach in which he assumed that the operation of a catchment on rainfall excess to produce an instantaneous unit hydrograph was analogous to the operation of routing an instantaneous inflow through a series of equal linear reservoirs. The assumption of equal linear reservoirs transforms the general equation of the unit hydrograph as follows

$$\frac{uT}{V_0} = \int_0^t \frac{\delta(t-\tau)}{\Pi(1+K_i D)} \cdot \omega(\tau')\, d\tau$$

$$= \int_0^t \frac{\delta(t-\tau)}{(1+KD)^n} \cdot \omega(\tau')\, d\tau$$

$$= \int_0^t \frac{1}{K} \cdot \frac{m^{n-1}e^{-m}}{(n-1)!} \cdot \omega(\tau')\, d\tau$$

where

$$m = \frac{t-\tau}{K}$$

Since Nash ignores the variation in translation time over the catchment, he assumes in effect that all points have the same translation time; that is, that the dimensionless time-area-concentration curve is a δ-function. This gives

$$\frac{uT}{V_0} = \int_0^t \frac{1}{K} \frac{m^{n-1}e^{-m}}{(n-1)!}\, \delta(0)\, d\tau$$

$$\frac{uK}{V_0} = \frac{m^{n-1}e^{-m}}{(n-1)!} \qquad m = \frac{t}{K}$$

which is Nash's solution for the instantaneous unit hydrograph.

Nash's solution has the advantage that the

TABLE 1

K/T	$u_{max}L/V_0$	
	Rectangle	Triangle
.25	.74	1.03
.50	.87	1.02
1.0	.95	1.01
2.0	.98	1.00
3.0	.99	1.00
4.0	1.00	1.00

S-hydrograph is given by

$$S(t) = \int_0^t u(0, t) \cdot dt$$

$$= \frac{V_0}{K} \int_0^t \frac{m^{n-1}e^{-m}}{(n-1)!} \cdot dt$$

$$= V_0 \int_0^t \frac{m^{n-1}e^{-m}}{(n-1)!} \cdot dm$$

$$= V_0 \frac{\int_0^t m^{n-1}e^{-m} \cdot dm}{(n-1)!}$$

$$= V_0 \cdot I\left(\frac{t}{K}, n-1\right)$$

where $I(\ )$ is the ratio of the incomplete to the complete gamma function, which has been tabulated by *Pearson* [1930]. Thus the ordinates of the finite duration unit hydrograph can be found from the difference between two tabulated values

$$u(D, t) = \frac{S(t) - S(t-D)}{D}$$

$$= \frac{V_0}{D}\left\{I\left(\frac{t}{K}, n-1\right) - I\left(\frac{t-D}{K}, n-1\right)\right\}$$

It is clear, therefore, that the solutions previously proposed as giving the form of the instantaneous unit hydrograph are special cases of the general equation derived earlier in the present paper. Even if the linear reservoirs are all assumed equal, the general method is extremely flexible, the result being dependent on the three numerical parameters T, K, and N, and on the functional parameters $\omega(\tau')$ and $n(\tau')$ defining the adjusted time-area-concentration curve and the reservoir

TABLE 2

Method	T	$\omega(\tau')$	K	N	$n(\tau')$	Degrees of freedom
Rational	variable	variable	0	0	—	2
Zoch-Clark	variable	variable	variable	1	at outlet	3
O'Kelly	variable	triangle	variable	1	at outlet	2
Nash	—	—	variable	variable	—	2
Using equal reservoirs	variable	variable	variable	variable	variable	5

distribution curve, respectively. The relation of the various special cases is shown in Table 2.

General solution using equal reservoirs

The assumption made in the Zoch-Clark method, that the reservoir storage in all parts of the catchment can be represented by a single linear reservoir at the outlet, seems unduly restrictive. Similarly, the neglect of translation in Nash's approach makes it impossible to analyze the effect on the unit hydrograph of such factors as the shape of the catchment and the ratio of overbank storage to channel storage. As indicated in the last section, the assumption that all the linear reservoirs in the idealized catchments are equal gives a solution having five degrees of freedom, compared with three and two degrees for the methods of Clark and Nash respectively. This additional flexibility is of great value, both for the direct analysis of the effect of catchment characteristics on the unit hydrograph, and also by providing additional parameters which can be used to improve the statistical analysis of derived unit hydrographs.

Any further generalization is probably not worthwhile. For the case of n unequal reservoirs, the solution of the general equation of the unit hydrograph is as follows

$$\frac{uT}{V_0} = \int_0^{t \leq T} \frac{\delta(t - \tau)}{\Pi(1 + K_i D)} \omega(\tau')\, d\tau$$

which gives

$$\frac{uT}{V_0} = \int_0^t \{C_1(\tau)e^{-(t-\tau)/K_1} + C_2(\tau)e^{-(t-\tau)/K_2} + \cdots + C_n(\tau)e^{-(t-\tau)/K_n}\} \cdot \omega(\tau') \cdot d\tau$$

The coefficients $C_1(\tau) \cdots C_n(\tau)$ are constant for the range corresponding to the reach between two adjacent reservoirs, but change on passing from downstream to upstream of each reservoir. Thus the computation is quite complex, becoming increasingly so as the number of reservoirs increases. Since this approach only gives more flexibility than the case of equal reservoirs if the number of unequal reservoirs is three or greater, the prospect is not encouraging. The case of unequal reservoirs is equally unpromising from the point of view of the numerical analysis of derived hydrographs. *Lanczos* [1957] has shown the uncertainty inherent in the analysis of exponential series. No advantage is obtained if two or more reservoirs are taken equal, since the number of terms remains the same and the individual terms are more complex. Simplification is only obtained if all the reservoirs are equal. From the practical viewpoint, the five degrees of freedom provided by the equal-reservoirs solution would appear to be ample for the purposes of both analysis and synthesis. In general, it may be concluded that the solution using equal reservoirs takes advantage of a considerable simplification without any appreciable loss of physical significance.

If all the linear reservoirs in the idealized catchment are taken equal in size, we have

$$\frac{uT}{V_0} = \int_0^t \frac{\delta(t - \tau)}{\Pi(1 + K_i D)} \cdot \omega(\tau') \cdot d\tau$$

$$= \int_0^t \frac{\delta(t - \tau)}{(1 + KD)^n} \cdot \omega(\tau') \cdot d\tau$$

$$= \int_0^{t/K} \frac{m^{n-1}e^{-m}}{(n-1)!} \cdot \omega(\tau') \cdot dm \qquad m = \frac{t - \tau}{K}$$

$$\frac{uT}{V_0} = \int_0^{t/K} P(m, n-1) \cdot \omega(\tau') \cdot dm$$

where

$$P(m, n - 1)$$

is the Poisson probability function. Since the latter function has been tabulated by *Pearson*

[1930] and *Molina* [1942], the computation of any ordinate of the instantaneous unit hydrograph by numerical integration is comparatively easy. For any ordinate (i.e., for a fixed value of t) the variables under the integral sign can be evaluated for each value of τ, and the products integrated using Simpson's Rule, or some equivalent. For any value of τ, $\omega(\tau)$ is given by the time-area-concentration curve, n by the distribution $n(\tau)$ of the N reservoirs each of delay time K, and m by the difference $(t - \tau)$. A convenient tabulation is shown in Table 3 where $K = IT$ and $N = 20$.

TABLE 3

t'	τ'	m	$n(\tau')$	$P(m, n-1)$	$\omega(\tau')$	Δu
1.2	.00	12	1	.006	0	.0000
	.05	11.5	1	.010	.2	.0000
	.05	11.5	2	.116	.2	.0002
	.10	11.0	2	.184	.4	.0007

It is important to remenber that each reservoir introduces a change in the value of n and hence a discontinuity in the value $P(m, n - 1)$. For this reason it is necessary to evaluate the integrand immediately upstream and downstream of each reservoir.

Application of proposed solution

The detailed discussion of possible applications of the proposed equal-reservoirs equation of the unit hydrograph is beyond the scope of this paper. Nevertheless, some indication is desirable of the relevance of the solution to key problems in theoretical and applied hydrology. Even if the shapes of the time-area-concentration curve and reservoir-distribution curve are kept constant, the method has three degrees of freedom, each characterized by a single numerical parameter of physical significance. Variation of the shapes of the two curves mentioned gives a tremendous additional flexibility, if required, either in analysis or statistical correlation.

One obvious application is the use of the solution to predict the effect of various catchment characters on the shape of the unit hydrograph. In this respect the method can do all that Clark's method can do, and all that Nash's method can do, as might be expected, since each of these have been shown to be a special case of the equal-reservoirs solution. In addition, the proposed equation can analyze the effect on the unit hydrograph of the distribution of storage within the catchment. Such an analysis is of more than academic interest, since it could be used to determine the relative effects of the various catchment characteristics. In this way the optimum number of parameters to be used in correlation studies or recorded data could be determined in advance. It is interesting to note that the unit hydrograph will approach the shape of the time-area-concentration curve under three conditions: (a) as K approaches zero, (b) as n approaches infinity, and (c) as the distribution approaches the case of uniform distribution. Analysis could indicate the relative importance and sensitivity of these factors and could assess the extent to which any one of them could replace the others without serious error.

Special cases of the general solution can be handled algebraically with interesting results. An example of this is the case treated by O'Kelly [1955], that is, a triangular time-area curve, $n = 1$, placed at outlet and two variable parameters. In this case, as the present author pointed out in the discussion of O'Kelly, the peak of the instantaneous unit hydrograph is described below

$$\frac{u_{\max} T}{V_0} = 4\left(1 - \frac{t_P}{T}\right)$$

$$e^{t_P/K} + 1 = 2e^{T/2K}$$

The latter equation can be expressed as

$$\frac{t_P}{K} = \log_e (2e^{T/2K} - 1)$$

which when expanded gives the following relationships within a few per cent

$$u_{\max} L = V_0$$

$$\frac{t_P}{L} = 1 - \left(\frac{K}{L}\right)^2$$

The same approach can be made for other special cases. Thus for Nash's assumption of n equal reservoirs and no variation in time of translation, we have

$$\frac{uK}{V_0} = \frac{m^{n-1} e^{-m}}{(n-1)!} \qquad m = \frac{t}{K}$$

This is a maximum when

$$(n-1)m^{n-2} - m^{n-1} = 0$$

i.e.

$$m_P = \frac{t_P}{K} = (n-1)$$

or

$$\frac{t_P}{L} = \frac{t_P}{nK} = \frac{n-1}{n}$$

$$\frac{u_{\max}K}{V_0} = P(n-1, n-1)$$

The above result is exact.

For the case of a rectangular catchment using Clark's method we have

$$t_P = T$$

$$\frac{u_{\max}T}{V_0} = 1 - e^{-T/K}$$

which is also an exact result.

These three special solutions are plosted in Figure 8. It is a simple matter to plot the coordinates of a derived hydrograph on this diagram for comparison with the three special cases.

It has been pointed out by Nash (private communication of draft paper) that the moments of the unit hydrograph are suitable as a basis for correlations in unit hydrograph analysis. Though this may not be rigorously true [*Kendall*, 1943], it is probable that such an approach is the best available at the present time. It is desirable, therefore, to seek a convenient expression for the moments of the general equation

$$\frac{uT}{V_0} = \int_0^{t/K} \frac{m^{n-1}e^{-m}}{(n-1)!} \cdot \omega(\tau') \cdot dm$$

which will be

$$M_R = \int_0^{\infty} dt\, t^R \int_0^{t/K} \frac{m^{n-1}e^{-m}}{(n-1)!}\, \omega(\tau') \cdot dm$$

Integration in the above form would appear to be difficult. However, if the moment is constructed by taking the contribution of each element of the catchment, integrating with respect to time, and then integrating over the catchment, a solution becomes possible. In this form the moment is

$$M_R = \int_0^T d\tau\, \omega(\tau') \int_\tau^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} \frac{dt}{K} \cdot t^R$$

$$m = \frac{t-\tau}{K}; \quad t = mK + \tau; \quad dt = K\, dm$$

so that

$$M_R = \int_0^T d\tau\, \omega(\tau') \int_0^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} (Km+\tau)^R \cdot dm$$

This can be evaluated by expanding the last term and taking advantage of the facts that τ is constant for the first integration and that

$$\int_0^{\infty} m^x e^{-m}\, dm = (x)!$$

Thus for the third moment we have

$$M_3 = \int_0^T d\tau\, \omega(\tau') \int_0^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} (Km+\tau)^3\, dm$$

expanding $(Km + \tau)^3$ as $K^3m^3 + 3K^2m^2\tau + 3Km\tau^2 + \tau^3$ and integrating with respect to m, term by term

$$\int_0^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} K^3m^3\, dm = K^3(n+2)(n+1)n$$

$$\int_0^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} 3K^2m^2\tau\, dm = 3K^2\tau(n+1)n$$

$$\int_0^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} 3Km\tau^2\, dm = 3K\tau^2 \cdot n$$

$$\int_0^{\infty} \frac{m^{n-1}e^{-m}}{(n-1)!} \tau^3\, dm = \tau^3$$

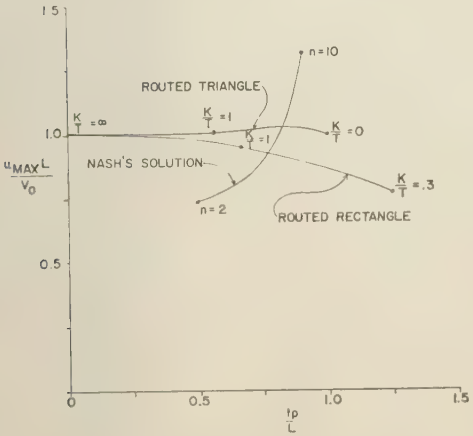


FIG. 8—Plotting of special solutions

so that

$$M_3 = \int_0^T d\tau\, \omega(\tau')\{K^3(n+2)(n+1)n + 3K^2\tau(n+1)n + 3K\tau^2 n + \tau^3\}$$

which can be expressed as

$$M_3(u) = K^3\{M_0(n^3\omega) + 3M_0(n^2\omega) + 2M_0(n\omega)\} + 3K^2\{M_1(n^2\omega) + M_1(n\omega)\} + 3K \cdot M_2(n\omega) + M_3(\omega)$$

Thus the third moment of the unit hydrograph can be expressed in terms of the moments of the distributions defined by products of the time-area curve $\omega(\tau')$ and the reservoir distribution curve $n(\tau')$. The other moments can be similarly found.

Errors of approximation in proposed solution

The solution proposed is applicable to a catchment consisting only of linear channels and linear reservoirs, in which the reservoirs are all equal and are so placed that there is the same number of reservoirs between the outlet and all points in the catchment having the same translation time. In this final section some computations are made to determine the order of the approximations involved in these limitations on the complete generality of the solution.

The separation of translation and reservoir action is fundamental to the approach made in the present paper. The assumption that any linear reach can be replaced by a number of linear channels and linear reservoirs placed in series is analogous to the lumping of parameters used in the analysis of electrical networks. It is reasonable to make this assumption at least in the first formulation of the problem. The representation of uniformly distributed storage as a form of translation is also reasonable. In natural catchments such storage might be of two types: (a) continuous overbank storage, or (b) a large number of patches of overbank storage each of which has a delay time small in comparison with the 'time of concentration' of the catchment. In the case of continuous overbank storage, the fact of linearity implies that the velocity is constant for all flows under consideration. Thus the delay time due to overbank storage is included in the translation time. The presence in a reach of a length of overbank storage is reflected in the idealized catchment by a length of linear channel whose translation time per unit length is greater than that of the remainder of the reach. In the case of a succession of patches of overbank storage, the assumption that this can be replaced by an equivalent translation channel is exactly true only when the number of patches is infinite. However, it is a good approximation in many cases. Figure 9

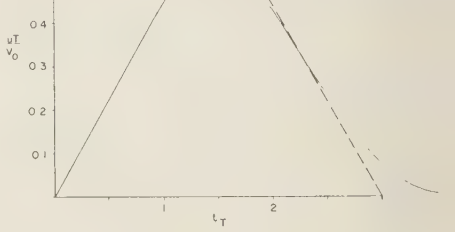


Fig 9—Translation approximated by twenty reservoirs

shows the comparison between the unit hydrograph for a triangular time-area curve with 20 reservoirs, each of delay time $T/10$, and the corresponding unit hydrograph for pure translation.

Once the assumption is made that translatory and reservoir action can be separated, the general equation of the unit hydrograph can be written as

$$u(0, t) = \int_0^{A(t)} \frac{\delta(t-\tau)}{\Pi(1+K_i D)}\, i\, dA$$

and the remaining assumptions are made only to obtain a convenient solution. The first of the two assumptions used in the present paper to produce a convenient solution is that the chain of reservoirs $K_1 \cdots K_n$ appropriate to any point is determined completely by the total translation time between the point in question and the outlet. For this condition the disposition of reservoirs along a tributary and along the main channel upstream of a junction will be the same. Many

natural catchments conform approximately to this condition of axial symmetry.

Computations were made to compare the total runoff due to each of three points of equal translation time, but with a different number of downstream reservoirs. Figure 10 shows the result for 8, 10, and 12 reservoirs, compared with the case for three branches each with 10 reservoirs. It will be seen from the figure that the case with the equal reservoirs gives a peak about 15 pct higher than that for the unequal reservoirs. If this difference occurs only over a small portion of the catchment, the corresponding error in the unit hydrograph will be much less than 15 pct. If differences occur over the greater part of the catchment, the error will be less than 15 pct in this case also, since the peak contributions from the different parts of the catchment will not coincide. We may conclude, therefore, that the above assumption, which transforms the surface integral of the general equation into an ordinary integral, will not give rise to serious error.

The remaining restriction on the generality of the proposed solution is the assumption that all the reservoirs are equal in size. Figure 11 shows the comparison between the outflow from a chain of three reservoirs of delay time K, and the outflow from a chain of reservoirs of delay time $0.5K$, K, and $1.5K$. Since the flows shown in Figure 11 are subjected to two smoothing processes (when integrating over the catchment and when passing to rainfall of finite duration),

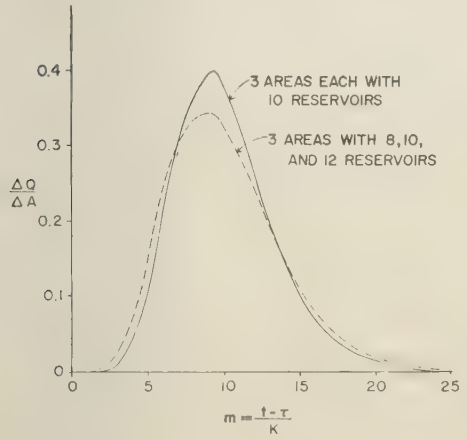


FIG. 10—Effect of unequal number of reservoirs

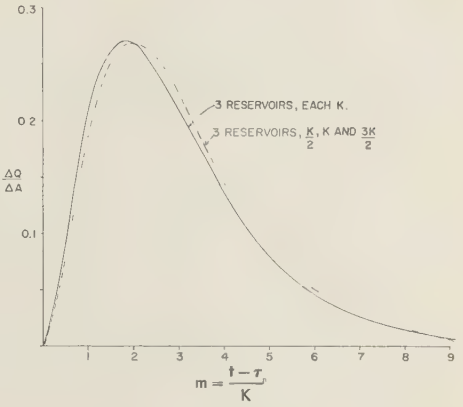


FIG. 11—Effect of unequal size of reservoirs

it will be obvious that the assumption of equal reservoirs will not be a source of appreciable error.

The trial computations made in this section show that for the cases examined the errors involved in the proposed solution are well within the normal accuracy of hydrological data. Accordingly, there is every reason to believe that the equation

$$\frac{uT}{V_0} = \int_0^t P(m, n-1)\omega(\tau')\, dm$$

$u = u(0, t)$ = ordinate of the instantaneous unit hydrograph
T = maximum translation time
V_0 = volume of rainfall excess
t = time elapsed since occurrence of rainfall excess
$P(m, n-1)$ = Poisson probability function
m = $(t-\tau)/K$ = dimensionless time factor
τ = translation time
K = size of linear reservoirs (all equal)
$n(\tau)$ = number of linear reservoirs downstream of τ
$\omega(\tau')$ = dimensionless time-area-concentration curve adjusted for variation in rainfall intensity

can be accepted as a flexible, convenient, and sufficiently accurate equation for the instantaneous unit hydrograph.

Acknowledgments—In the development of his ideas on the subject of the unit hydrograph, the author was helped greatly by conversations over a number of years with the members of the hydrology group of the Irish Office of Public Works, the late

J. J. O'Kelly, J. P. Farrell, and J. E. Nash. Comments by J. W. Howe and M. R. Bottaccini, of the Department of Mechanics and Hydraulics at the State University of Iowa, were also of assistance. G. P. Foley, of Civil Works Department, Electricity Supply Board, assisted in the computations.

References

Clark, C. O., Storage and the unit hydrograph, *Trans. Am. Soc. Civil Engrs.*, **110**, 1419–1446, 1945.

Dooge, J. C. I., Synthetic unit hydrographs based on triangular inflow, M. S. thesis, State University of Iowa, 1956.

Dooge, J. C. I., The rational method for estimating flood peaks—Irish contributions to the technique, *Engineering*, pp. 311–313, Sept. 6, 1957, pp. 374–377, Sept. 20, 1957.

Horton, R. E., Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology, *Bull. Geol. Soc. Amer.*, **56**, 275–370, 1945.

Kendall, M. G., *The advanced theory of statistics*, **1**, chap. 3 and 4, C. Griffin & Co., London, 1943.

Lanczos, C., *Applied analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1957.

Molina, E. C., *Poisson's exponential binomial limit*, D. Van Nostrand Co., New York, 1942.

Nash, J. E., The form of the instantaneous unit hydrograph, *C. R. et Rapports, Assn. Internat. Hydrol. IUGG*, Toronto, 1957.

O'Kelly, J. J., The employment of unit hydrographs to determine the flows of Irish arterial drainage channels, *J. Inst. Civil Engrs. (London)*, **4**, 365–445, 1955.

Pearson, K., Tables of the incomplete Γ function (1922) and Tables for statisticians and biometricians, *Biometrika*, University College, London, 1930.

Sherman, L. K., Streamflow from rainfall by the unit hydrograph method, *Eng. News-Record*, **108**, 501–505, 1932.

Zoch, R. T., On the relation between rainfall and stream flow, *Monthly Weather Rev.*, **62**, 315–322, 1934; **64**, 105–121, 1936; **65**, 135–147, 1937.

(Manuscript received June 11, 1958.)

Measuring Soil Moisture over Large Areas with Single Installations of Moisture Units

JOHN L. THAMES

Southern Forest Experiment Station
Forest Service
United States Department of Agriculture
New Orleans, Louisiana

Abstract—It seems possible that, with proper calibration, single installations of electrical soil-moisture units can be used to estimate moisture contents of outlying sites receiving similar amounts of rain but differing in vegetation and soils. In tests with data from northern Wisconsin and west-central Mississippi, predicted moisture contents agreed closely with those obtained by on-site measurements. Results were most accurate at seasons when the soils remained near the wilting point or near field capacity and were least accurate during periods of recharge.

In the area concept of soil-moisture measurement, readings from a field-calibrated moisture unit in a given soil layer are considered an index of the average moisture content of that layer over the calibration area [*Lull* and *Reinhart*, 1955; *Olson* and *Hoover*, 1954]. Usually calibration areas extend only a few feet around the unit installation.

This paper proposes an application of the area concept to estimate soil moisture over larger areas not necessarily uniform in soil or cover.

Examination of soil-moisture records at the Vicksburg Research Center* of the Southern Forest Experiment Station indicated that, within an area of nearly equal rainfall, the wetting and drying rates on different sites might be similar even though absolute values of saturation and wilting point differed. It followed that, if the relationship of drying and wetting rates between sites could be established, a single tier of units might be used to indicate soil moisture at widely separated sites, provided they had complete plant cover and were exposed to the same weather.

The concept was tested at Vicksburg, Mississippi, and Rhinelander, Wisconsin, mainly with data previously gathered in the course of soil trafficability studies by the Vicksburg Center.

Application of any method for estimating soil moisture at one location from measurements made at a distant point must assume proportionate rates of soil-moisture loss throughout the working range of moisture for the soils at both locations. *Thornthwaite* [1944], in applying equations for calculating potential evapotranspiration, implied that rates of soil-moisture loss are independent of the character of plant cover and of soil type, provided that the different types of vegetation have the same seasonal use of moisture and the same root depth. This was borne out by experience at Vicksburg [*U. S. Army Corps of Engineers*, 1954], where it was found that within the root zone in the top foot of soil, differences in vegetative cover were not reflected by corresponding differences in soil-moisture depletion, even when moisture was limiting.

Mather [1950] states that evapotranspiration is as much an element of climate as is precipitation. *Penman* [1946] has reasoned that, if meteorological conditions are equal, potential evapotranspiration from two vegetated areas will be similar and more a function of climatic factors than of soil and vegetation types.

Study areas—Four groups of sites were studied, each group comprised of sites with similar rainfall. Soil properties are summarized in Table 1.

The sites of Groups 1 and 2, near Rhinelander, Wisconsin, were on glaciated soils. The four sites in Group 1—Sortek-1, Gross-1, Peterson, and Timber Lake—had mixed grass and herbaceous cover. The three sites of Group 2—

* Maintained at Vicksburg, Mississippi, by the Southern Forest Experiment Station, Forest Service, U. S. Department of Agriculture, in cooperation with the Waterways Experiment Station, Corps of Engineers, U. S. Army.

TABLE 1—*Soil properties*

Site identification	Soil series	Depth of measurement (inches)	Mechanical analysis, by weight (per cent)			Bulk density (g/cc)	Soil moisture, per cent by volume, at pressures of		
			Sand	Silt	Clay		.005 atmosphere	.060 atmosphere	15 atmospheres[1]
RHINELANDER, WISCONSIN									
Group 1									
Sortek-1	Spencer	0–12	28	65	7	1.27	48.3	43.2	12.7
Gross-1	Spencer	0–12	37	58	5	1.15	41.4	38.0	10.4
Peterson	Antigo	0–12	36	59	5	1.40	42.0	37.8	7.0
Timber Lake	Vilas	0–12	83	15	2	1.51	40.8	33.2	6.0
Group 2									
Sortek-2	Spencer	0–12	23	71	6	.96	41.3	36.5	7.7
Gross-2	Spencer	0–12	26	68	6	1.06	41.3	36.0	8.5
Gudis	Antigo	0–12	45	50	5	1.20	44.4	38.0	9.6
VICKSBURG, MISSISSIPPI									
Group 3									
Hardwood	Memphis	0–12	6	80	14	1.12	39.2	35.8	7.8
		12–24	4	80	16	1.28	40.0	37.1	7.2
Pine	Memphis	0–12	7	69	24	1.30	42.9	39.0	14.3
		12–24	6	72	22	1.35	43.1	40.6	13.5
Durden	Briensburg	0–12	2	82	16	1.29	46.4	42.5	12.9
		12–24	4	79	17	1.32	47.5	40.9	10.6
Group 4									
Park-1	Loring	0–12	7	74	19	1.38	44.1	35.9	13.8
		12–24	4	71	25	1.40	46.3	38.2	13.6
Park-2	Loring	0–12	7	74	19	1.29	42.6	33.5	12.9
		12–24	5	73	22	1.36	45.2	36.7	14.1

Sortek-2, Gross-2, and Gudis—were in 40- to 50-year-old stands of aspen. Soils of the Sortek and Gross sites were developed on level topography from till plain materials derived from crystalline rocks. Peterson and Gudis were upland terrace sites on soils developed from glacial alluvium. The Timber Lake site was on an upper slope in gently rolling topography on soil developed from sandy drift.

The other site groups were at Vicksburg, Mississippi, on loessial soils. Group 3 included three dissimilar sites—Hardwood, Pine, and Durden. The Hardwood site, on an upland ridge, was forested with native hardwood species 50 to 74 years old. As nearly as could be determined, the soil had never been cultivated. The Pine site occupied the crest of a slope that had been severely eroded into the *B*-horizon and supported 15-year-old pines. Durden, a site in a nearly level stream bottom on alluvial soil, had a cover of annual weeds and grasses. The Group 4 sites, Park-1 and Park-2, occupied mid-slope positions on rolling topography and supported perennial grasses.

Measurements and analysis—The records for the Wisconsin sites extended from May 1 through December 5, 1953. King-tube samples had been collected weekly from the 0- to 3-inch, 3- to 6-inch, 6- to 9-inch, and 9- to 12-inch layers from four 6- by 6-foot sampling plots at each of the sites. At Sortek-1 and Sortek-2 fiberglas soil-moisture units had been installed adjacent to the sampling plots and read daily. At the other sites additional gravimetric samples were taken at intervals from the 0- to 6-inch and 6- to 12-inch layers.

All Vicksburg sites had fiberglas moisture units. Daily readings were on record for the Hardwood, Pine, and Durden sites from April, 1952 to March, 1953, and for the two Park sites from April, 1951 to March, 1952. Successive depths from 3 inches to 24 inches were represented. At 15 to 20 times on each site, soil cores had been taken from a 6- by 6-foot sample plot adjacent to the unit installation.

Soil texture, bulk density, and soil-moisture tension values had been determined for all sites and all depths sampled. Gravimetric determina-

TABLE 2—*Mean differences of sample points from calibration curves*

Site identification	Distance from reference tier (air miles)	Mean differences (per cent by volume) at depths of								No. of comparisons
		0–3 in.	3–6 in.	6–9 in.	9–12 in.	12–15 in.	15–18 in.	18–21 in.	21–24 in.	
RHINELANDER, WISCONSIN										
Group 1										
Sortek-1	—	2.3	3.1	3.0	4.0	—	—	—	—	27
Gross-1	0.30	3.2	2.0	2.7	3.0	—	—	—	—	26
Peterson	1.06	2.3	2.2	2.4	2.7	—	—	—	—	22
Timber Lake	1.05	3.0	1.3	2.3	2.5	—	—	—	—	26
Group 2										
Sortek-2	—	2.9	2.0	3.5	3.0	—	—	—	—	28
Gross-2	.30	3.1	2.0	1.3	1.7	—	—	—	—	26
Gudis	1.18	5.0	2.7	3.6	3.1	—	—	—	—	23
VICKSBURG, MISSISSIPPI										
Group 3										
Hardwood	—	1.3	1.7	2.0	2.0	0.7	2.3	1.3	1.1	15
Pine	.30	2.0	2.1	1.8	1.3	2.3	2.0	1.0	2.0	16
Durden	.32	1.3	2.3	1.7	1.4	1.3	1.6	1.3	1.9	20
Group 4										
Park-1	—	1.7	1.1	1.0	2.2	2.6	1.8	2.4	2.0	18
Park-2	.04	1.3	2.0	1.2	1.5	1.7	1.3	1.8	1.7	16

tions were on an oven-dry basis. Readings from the fiberglas units were converted to the logarithms of the resistances in ohms at 60°F.

To test the practicability of the method, a single tier of moisture units within each group of sites was used as the reference location. Moisture contents of gravimetric samples from each site in each group were plotted against readings at this tier. Calibration curves were prepared for the soil layers of each site by the methods described by *Reinhart* [1953]. From daily resistance records at the reference tier, retrospective daily records of soil moisture were then constructed for each outlying site. Distances between reference and outlying sites ranged from 0.04 to 1.18 miles (Table 2).

Units of the Sortek-1 and Sortek-2 reference installations were calibrated with the gravimetric field samples from each site in the respective groups. The units of the Hardwood and the Park-1 sites were taken as the reference tiers for the sites in Groups 3 and 4.

Results and discussion—For the sites and depths studied, there proved to be no significant differences between soil moisture content as determined from a reference tier and that measured directly at the outlying sites.

Preliminary comparisons were made by calculating for each day of sampling the differences between the gravimetric data and the values read from the calibration curves. To facilitate computation, soil-moisture contents of all depths at each site were totaled. Average differences from the curves for the sites adjacent to the reference tiers were then compared with the differences from the curves for the outlying sites within each of the four groups. A '*t*' test of the pooled variances indicated that differences were not significant.

At Vicksburg, units had been installed and calibrated at the outlying sites as well as at the reference sites. Here, deviations of gravimetric samples for the original calibration curves were compared with the deviations of the same samples from curves based on the reference tiers. Again a '*t*' test failed to show significance.

In Figure 1, representative curves for the Pine and Durden sites, based on readings from the Hardwood reference tier, are compared with the original calibration curves. Data for an additional site, Rifle, are included to show the lack of relationship between sites with different rainfall. The Rifle site was 2.9 miles from the Hardwood site. In soil and plant cover

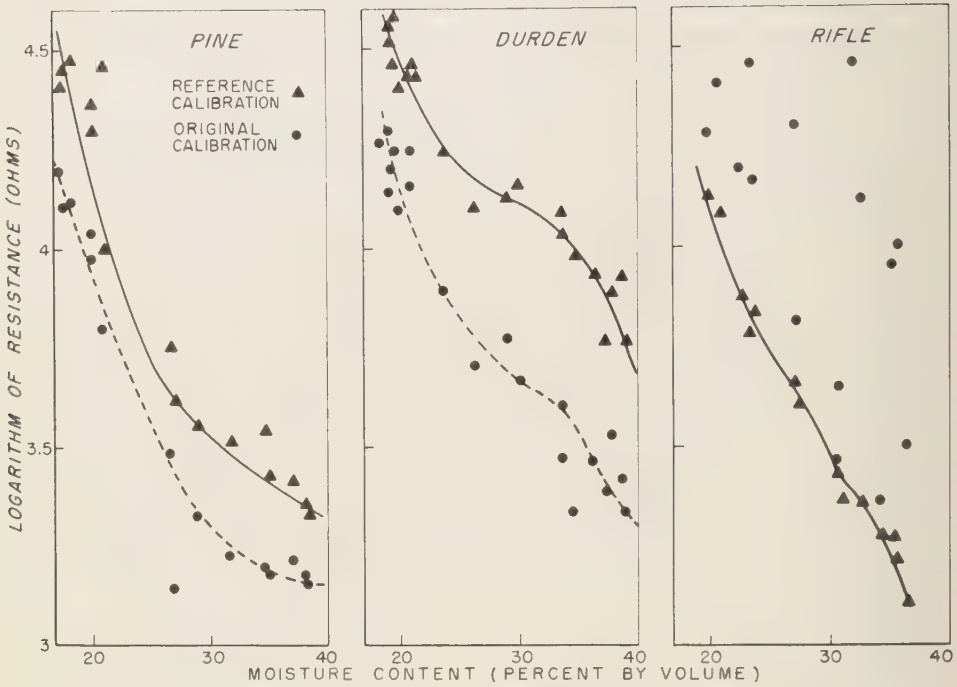


FIG. 1—Sample calibration curves for Vicksburg sites, 21- to 24-inch depth

it was similar to Durden, but some individual rainstorms differed as much as 0.55 inch between Rifle and the Group-3 sites.

An obvious feature of the data for sites with the same rainfall is that the curves do not coincide. This is true between sites—i.e., where different gravimetric samples from different sites have been related to resistances read at the reference tier. It also holds within sites—where single gravimetric samples from one site have been related to resistances read both at the reference tier and at the outlying tier. The shape and graphical position of the curves compensate to a great extent for differences in the physical properties of the soils and electrical characteristics of individual moisture units. Mean differences of field samples from the curves for all sites are given in Table 2.

Because the calibration curves were drawn freehand, the comparisons just made are not exact. However, they do indicate the relationship between the soil-moisture regimes of the sites in each group.

A more reliable test was made for each outlying site by comparing moisture contents as predicted from the reference-unit readings with soil moisture measured independently by standard methods. At Vicksburg the standard of comparison was the set of records originally made with the moisture units on the outlying sites. At Rhinelander gravimetric sampling was done expressly to check the method. Results at both locations are given in Table 3.

The high correlations seem partly due to similarity in the moisture characteristics of the soils. In general, the lower correlations are associated with correspondingly greater differences in tension values and water-yield characteristics. Timber Lake, with lowest correlation, differed in soil-moisture content from Sortek-1 by 7.5 pct at field capacity and by 6.7 pct at wilting point.

The generally lower correlations for the Rhinelander sites may be attributed to the greater variations common to glaciated soils. In a study reported by *Lull* and *Reinhart* [1955], variation of moisture sampling within a 0.92-acre area was computed for several soils. Average standard deviation for the 0- to 6-inch layer in the dry ranges of soil moisture was 2.1 pct for a

TABLE 3—*Comparison of predicted and actual moisture contents*

Site identification	No. of comparisons	Mean differences (pct by vol.)	Standard error of estimate (pct by vol.)	Coefficient of correlation
RHINELANDER, WISCONSIN				
Group 1				
Gross-1	8	1.4	1.7	0.83
Peterson	9	1.3	.8	.89
Timber Lake	9	1.4	2.3	.77
Group 2				
Gross-2	8	1.5	1.6	.84
Gudis	8	1.2	1.8	.87
VICKSBURG, MISSISSIPPI				
Group 3				
Pine	312	.5	.7	.92
Durden	294	.7	.9	.88
Group 4				
Park-2	253	.6	.7	.94

Memphis silt loam at Vicksburg and 4.3 pct for a Spencer silt loam at Rhinelander.

Figure 2 shows the on-site soil moisture record and the record determined from the reference unit installation for the Pine site at Vicksburg. Most of the differences between the two records can be accounted for by unequal wetting due to differences in soil and microtopography between sites. Rains that satisfy the available storage space of the soils, such as those occurring from January through March, present no problem. After such rains the soils at both locations approached field maximum at the same time and dried out proportionately between storms.

High-intensity storms that exceed maximum infiltration rates and storms where total rainfall is less than the available storage space of the soils may result in unequal wetting. When this occurs, the tendency for soil moisture held at lower tensions to be depleted faster than that held at higher tensions helps to reduce inequalities by bringing tension values and moisture contents together. In Figure 2 the effects of this governing relationship show up in the depletion periods of May 29, August 8, and August 21.

Greatest differences between the two records occurred during the soil moisture recharge period, which in the present study occurred during November and December. To some extent these differences may have been due to the erratic behavior of moisture units during wetting, as noted by *Horton* [1955].

Summary—Results of this study indicate that single installations of soil-moisture units may be used to estimate moisture content of outlying areas where rainfall amounts are nearly equal but vegetation and soils may vary.

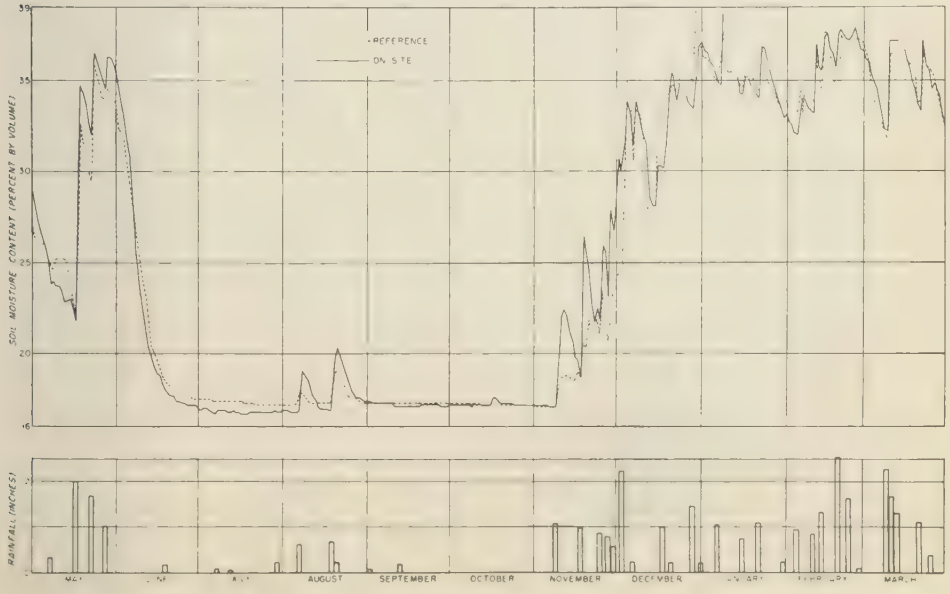


FIG. 2—On-site and reference soil moisture records, Pine site, Vicksburg, 1952-1953

The method was tested with data from sites in northern Wisconsin and west-central Mississippi. Readings on moisture units at central sites were related to moisture contents determined by gravimetric sampling on other sites up to 1.18 miles distant. With the resulting calibration curves, daily readings at the central moisture units were used to predict moisture contents of the outlying sites.

For the conditions studied, results thus secured were closely correlated with those obtained by on-site measurement. The method worked best in the summer, when the soils remained near the wilting point, and in portions of the winter and spring when frequent rains kept soils near field capacity. Accuracy was fair during drying periods but poor during periods of recharge.

References

Horton, J. S., Use of electrical soil-moisture units in mountain soils, *Proc. Western Snow Conf.*, **23**, 20–26, 1955.

Lull, H. W., and K. G. Reinhart, Soil moisture measurement, *U. S. Forest Serv. South. Forest Expt. Sta. Occas. Paper 140*, 56 pp., 1955.

Mather, J. R., Manual of evapotranspiration, *Johns Hopkins Univ. Laboratory of Climatology: Micrometeorology of the surface layer of the atmosphere, Supplement to Inter. Rep. 10*, April 1 to June 30, 1950.

Olson, D. F., Jr., and M. D. Hoover, Methods of soil moisture determination under field conditions, *U. S. Forest Serv. Southeastern Forest Expt. Sta., Sta. Paper 38*, 28 pp., 1954.

Penman, H. L., Some aspects of evaporation in nature, *Sci. J. Roy. Coll. Sci.*, **16**, 117–129, 1946.

Reinhart, K. G., Installation and field calibration of fiberglas soil-moisture units, *U. S. Forest Serv. South. Forest Expt. Sta. Occas. Paper 128*, pp. 40–48, 1953.

Thornthwaite, C. W., Report of the committee on transpiration and evaporation, 1943–44, *Trans. Amer. Geophys. Union*, **25**, 686–693, 1944.

U. S. Army Corps of Engineers, The development of methods for predicting soil moisture content, *Progress Rpt. 3, Forecasting trafficability of soils, Technical memorandum no. 3-331*, Waterways Expt. Sta., Vicksburg, Miss., 1954.

(Manuscript received August 11, 1958.)

Particle Coatings Affecting the Wettability of Soils*

Bessel D. van't Woudt

Agricultural Engineering Department
University of Hawaii, Honolulu

Abstract—The fundamental principles underlying the process of wetting show that a reduction in the surface tension of a solid substance to be wetted reduces the wettability. Conversely, a reduction in the surface tension of the applied liquid increases the wettability. The wetting of a soil, which exhibits water-repellency in an air-dry state, can sometimes be achieved by leaving water in contact with this soil. There is some evidence that this wetting is achieved by an interaction of the solid and the liquid phase, leading to a reduction in the surface tension of the liquid. The phenomenon of advancing and receding angle of contact had led to an hypothesis that the angle of contact changes as the soil wets up or dries out. Some evidence is presented which does not support such an assumption, as the angle of contact seems fixed at a moisture content just above air-dry. There is some evidence that under certain conditions an improvement in the base status of the soil may lead to improved wettability. Particle coatings by hydrophobic films are in some cases responsible for the observed water-repellency. There is some evidence that this film is strongly adsorbed. No quantitative relationship could be detected between the characteristics of substances extracted by chloroform from a problem soil and the degree of unwettability of such a soil.

Introduction—At the last meeting of the Regional Research Committee of the seven Western States the desirability was discussed of initiating a research program on the effect of particle coatings on (1) the wettability of soils and (2) the movement of water through soils. It had been tempting to assume that the generally observed relationship between soil permeability and soil moisture content is related to a change in the angle of contact between water and soil at varying moisture contents. The assumption was that the angle of contact increases as the soil dries out, thus creating increased resistance to flow. Conversely, one could assume that as the soil moisture content increases the angle of contact decreases, thus permitting higher permeability. Some impetus to this thinking had been given by the reports made from time to time on the unwettability of dried-out soils. The following observations may make some contribution towards a solution of this problem.

The theory of wetting—The work of adhesion performed when a liquid is brought into contact with a solid is expressed by the equations:

(a) $W_{sl} = t_l(1 + \cos\theta)$ (Young's equation) (1)

and

(b) $W_{sl} = t_s + t_l - t_{sl}$ (Dupré's equation) (2)

where W_{sl} is the work of adhesion, t_l is the surface tension of the liquid, t_s is the surface tension of the solid, and t_{sl} is the interfacial tension between solid and liquid. On combining (1) and (2) it is found that

$$t_s + t_l - t_{sl} = t_l + t_l \cos\theta \qquad (3)$$

and

$$t_s - t_{sl} = t_l \cos\theta \qquad (4)$$

so that

$$\cos\theta = \frac{t_s - t_{sl}}{t_l} \qquad (5)$$

Any quantities which produce a quotient in excess of one will cause the angle of contact, $\cos\theta$, to be zero. If the denominator exceeds the numerator in value, the quotient is less than one, and a definite angle of contact is formed. The angle of contact varies from zero to 90 degrees when solids are wettable by a given liquid, the rate of wetting depending upon the actual angle. At a 90-degree angle the self-attraction of the liquid molecules is equal to the inter-attraction of the liquid to the solid molecules. Above 90 degrees the self-attraction of the liquid molecules exceeds the liquid-solid attraction. At a contact angle in excess of 90 degrees, liquid does not spread over a solid surface but collects in rounded droplets exhibiting a characteristic angle of contact.

* Published with the approval of the Director of the Hawaii Agricultural Experiment Station as Miscellaneous Paper No. 107.

As is seen from equation (5), a high surface tension of a solid enhances the possibility of the spreading of a liquid, and conversely a low surface tension of a solid, as is exhibited by certain soft substances, such as waxes and resins [*Bartell* and *Zuidema*, 1936] tends to restrict wetting. Also, as one would expect from the above relationships, organic liquids with a surface tension of 20 to 30 dynes per centimeter at room temperature may wet a solid which is not wetted by water with a surface tension of about 73 d/cm at room temperature.

Water repellancy in soils—A resistance to wetting is typically exhibited by dried-out peat and muck soils and by sandy soils under coniferous or heath growth. According to *Zuncker* [1930], unwettability of soil is observed under such conditions whenever the base status is so low that organic acid radicals, considered to have hydrophobic properties, are not balanced by bases. However, by necessity, the negatively-charged radicals need to be balanced by positively-charged ones. In a soil of low base status this could be achieved by hydrogen or weakly basic cationic complexes. For Zuncker's hypothesis to hold true it will be necessary that the hydrophobic characteristics of the acid radicals are not offset by the balancing cationic complexes. At this stage such a mechanism is not understood.

Under certain conditions a resistance to wetting is probably also produced by the formation of water-repellent calcium or magnesium soaps formed on application of calcium or magnesium fertilizers. Evidence for this was presented by *Wander* [1949] for sandy soils under citrus trees in Florida. Fatty acids dissolved from this soil (which was unwettable in a dry state) and added to another wettable soil did not produce unwettability in the second sample unless calcium or magnesium hydroxide were added. *Domingo* [1950] quoted some inaccessible European literature in which iron was also qualitatively related to a resistance to wetting in soils. The possibility of lack of intimate contact between a liquid and a solid surface as a result of air films seems to apply to special cases, notably where air is locked in the voids of powdery or granular material.

Wetting behavior—Use was made of various soil phases found on a slope 50 feet long in volcanic ash soil, relatively unweathered Taupe ash of a rhyolitic nature, on the central plateau in the north island of New Zealand. These soil phases exhibit varying degrees of water repellency as was systematically determined on 10-gram air-dry samples placed on the surface of water in observation dishes. The samples were taken from the *A*, *B*, and *C* horizons from five successive locations on the slope and one position at the foot of the slope. Interestingly, an increasing degree of water repellency was apparent in passing from the samples of highest base status [*van't Woudt*, 1955], the *A* horizon at the foot of the slope, to those of lowest base status, the *C* horizon in the root zone below the top of the slope. With the exception of the latter ones, all the samples could be wetted by leaving them in contact with water for varying periods.

The phenomenon of the 'contact period' has been noted by several workers, but no satisfactory explanation has been offered. The following observations seem to have a bearing on the problem. If soil requiring a contact period for wetting is placed on a watchglass under a binocular microscope with incident light and a drop of water is placed on its surface, it can be seen that the drop rests initially on the soil surface with a high angle of contact. At this time the surface of the drop is clear and reflectant. Immediately after contact between soil and water some movement of fine soil and humus particles takes place at the interface, whereby some particles are bodily moved in position, and turned over. Some of the particles of elongated shape actually protrude into the drop during this process. Using soil which can be wetted by leaving it in contact with water for a few minutes, this process is observed to take place during about two minutes. After that time the view through the drop becomes obscured because of a covering of the outside of the drop by a fine film of soil and humus particles which creep up evenly from all sides towards the center. Just preceding the formation of this clearly visible film of fine particles, a much thinner film is formed of such minute particles that they cannot be distinguished individually through the microscope. Its presence can be detected only by the fact that the surface of the water drop changes from reflectent to dull. At the moment the film of the larger particles converges in the center on top of the drop, the shape of the drop starts a gradual change which occurs during

the next 60 seconds. The drop gradually flattens and it can be seen that this is associated with a change in contact angle from a high to a low one. As the angle of contact approaches zero, the drop on the soil surface suddenly collapses, leaving a wetted spot on the soil surface. The wetting process thus takes three to four minutes in this case. Using the same soil and allowing a drop of water to be dropped from a height of eight inches, for example, the first stage of the process is apparently eliminated or proceeds at such a rate that it cannot be observed under the microscope. The drop becomes immediately covered by a film of particles. Also in this case the angle of contact is immediately lower and, consequently, the drop is flatter. In this case the wetting process takes from one to two minutes.

If soil with a greater resistance to wetting is used, several variations in the above pattern can be observed. In some cases it takes as much as ten hours for the surface film of fine particles to close in at the top of the drop of water. During the first two minutes after the application of the drop a faint agitation of the soil and humus particles can be observed at the soil and water interface. Then the agitation ceases, but the very slow advance of the film proceeds.

One explanation for these observations is that the film formation occurs because of an establishment of contact between water and hydrophilic loci on the film-forming particles. Presumably once the complete film is formed the surface tension of the water drop is reduced, after which the changed relationship $(t_s - t_{sl})/t_l$ favors a lower angle of contact. The following observations seem relevant. The impact of a drop of water allowed to fall from some height, as occurs during rain, might lower momentarily the intramolecular attraction in the drop, leading to a reduction in its surface tension. Increasing the temperature of the applied water reduces the surface tension, and in line with this the time required for wetting is reduced, the actual reduction in time increasing with increase in water temperature. Wetting can be accelerated by stirring water into the soil. This would create an opportunity for contact between hydrophilic loci and water; whether this would also lead to a reduction in surface tension of the water would be hard to prove.

The advancing and receding angle of contact—The fact that substances are hard to wet when dry, but readily wettable at a low initial moisture content is widely observed in laundry work. As far as is known to the writer, no satisfactory explanation for this observation has been presented as yet. Some workers have considered that this phenomenon is related to the overturning of hydrophilic or hydrophobic bonds at the surface of the solid substance [e.g., *Adams*, 1948]. A further explanation is suggested by the foregoing observations on wetting behavior as the surface tension of the liquid is seemingly lowered during a 'contact period.' Further work along the above lines is required; for the time being it is of interest to study the implications of the advancing and receding angle of contact as far as moisture behavior in soils is concerned.

This was studied in small-scale models, using dune sand and volcanic ash as soils, both of which showed a moderate degree of resistance to wetting in an air-dry state. The soil was shaped in trays, measuring 12 × 15 inches, into microrelief with slopes of varying degrees, about 10 inches long. The soil was wetted by overhead spraying of water by an insecticide spray pump in a wind-free laboratory. Results for air-dry soils were described by *van't Woudt* [1954] where it was shown that the resistance to wetting led to subsurface storm flow, resulting in deep wetting at the foot of the slope and shallow wetting on the slope itself. These experiments were later extended to soil with varying degrees of moisture content. An initially desired moisture content was created by kneading known quantities of water into known quantities of soil, assuring by sampling for moisture content that a reasonably even moisture distribution had been obtained at the end of this operation.

In an air-dry soil the advance of a 'wet front' can readily be detected, but in moist soil visual observations become unreliable. A number of techniques were developed, which in sequence show refinement. After water was applied, as described, to a moist soil, samples were collected along an exposed transect cut by knife in the direction of the slope, systematically taking samples from successive cubic inches of soil in a vertical and horizontal direction. This was followed by applying water colored by indigo carmine, a dye which was selected out of 20 dyes and chemicals tried for this purpose. The problem encountered was that, in order to be

soluble in water, the compound had to have polar bonds; however, the polarity of such substances caused them in turn to be adsorbed on the soil particles. Indigo carmine showed the least adsorption; some of the dye followed the water down into the soil, but color variation was difficult to detect. Blocks of soil were then cut away in a fashion similar to that described for the micro-soil sampling method, and the blocks of soil were immersed in water in beakers. After leaving the soil to stand overnight, the supernatant liquid was decanted off and the degree of blue coloration of the water, caused by redissolution of the dye, was plotted. The last method used, which gave the most positive results, consisted of exposing a transect, as before, and applying heat from an electric room heater at close range. By this method the pattern of moisture distribution was clearly shown, as the soil unwetted by the water application dried out first. All of these methods showed that a resistance to wetting did not have any effect on moisture movement at a moisture content above that of air dry, and, therefore, that there was no variation in contact angle. The evidence for this is as follows. It was demonstrated by *van't Woudt* [1954] that a difference between the angle of contact in a wetted topsoil layer and that in the underlying air-dry soil led to lateral water movement through the wetted top layer. If a difference in contact angle had persisted at a moisture content above air dry, lateral water movement would similarly have been demonstrated. While about ten different tests for moisture content were made, in none of the cases could a trace of lateral movement be demonstrated. Consequently, according to these observations, the angle of contact in a wet top layer of soil and an underlying moist soil was the same.

TABLE 1—*Contents of chloroform-soluble constituents and their chemical characteristics in Taupe volcanic ash soil.*

The samples are arranged from top to bottom in order of increasing degree of water repellency. Determination on samples of 2000 gram

Soil description	Dissolved matter as a dry weight percentage (a)	Melting point degree F	Saponification value	Percentage unsapon. matter in (a) (b)	Iodine value of (a)	Iodine value of (b)
Soil under sod	.4		165.6	11.6	34.5	60.2
Soil A_1 layer under conifers	.5	60–67	229.3	44.2	62.1	71.1
Soil A_1 layer under heath	.7		325.4	13.0	35.3	64.1
Soil A_1 layer under heath	1.0		248.0	16.9	37.5	64.1
Soil A_1 layer under heath	1.1		201.0	18.5	43.9	74.8
Soil B layer under conifers	.2	40–75	179.4	41.2	72.6	82.2
Soil A_1 layer under conifers	.5		174.8	12.4	36.7	75.2
Soil A_1 layer under conifers	.4		169.1	14.8	43.0	59.5
Soil A_2 layer under conifers	.16		120.0	47.6	69.8	64.0
Monoao heath, buds and young wood	6.9					
Monoao heath, basal stems	6.0					

Analysis by Fat Research Bureau, Dept. Sci. and Ind. Res., Wellington, N. Z.

These observations are important in that they do not support a hypothesis that the angle of contact changes throughout the range of moisture contents found in a field soil. Therefore, they do not support a view that an increase in permeability with increase in soil moisture content is related to a lowering of the angle of contact.

Base saturation and unwettability—Zuncker's [1930] observations that soils show a resistance to wetting under certain types of vegetation and when the soil has a low base status may be of considerable practical interest. In volcanic ash soils water-repellent soils are commonly met with under native heath vegetation; under nearby coniferous trees a resistance to wetting is also marked, but on intervening sites where the soil has been under introduced grasses for 20 years or more no resistance to wetting is observed. The dominant shrub of the heath vegetation is monoao, known for its high oil content which makes it burn with a smoky flame. Some relevant analytical data on the content of chloroform-soluble substances are shown in Table 1. The interest centers here around the observation that if soil samples taken from underneath monoao are given various quantities of lime, the degree of unwettability decreases gradually as the lime application is increased; water-repellency is absent where lime applications of five tons per acre are worked into six inches of topsoil. In line with this, the use of a $N/100$ ammonia solution greatly accelerates wetting and a $N/1$ solution eliminates any water-repellency. (The ammonia would create an immediate base saturation by NH_4^+ ions.) These observations support the field evidence that a resistance to wetting in soil diminishes as heathland is developed into pasture which receives topdressing. This in turn probably leads to an improvement in the base status in this particular volcanic ash soil.

Particle coatings—It has been demonstrated by several workers that if difficultly wettable soil is extracted with ether, alcohol, methanol, or other organic solvents, the extract thus obtained added to a readily wettable soil makes this soil exhibit a resistance to wetting in an air-dry state. Volcanic ash soil extracted with ether for 20 hours in a Soxhlet apparatus still remained unwettable after that period. This indicates that the soil particles are not coated by a loosely adhering film, but that the hydrophobic properties are imparted by hydrophobic bonds on complex radicals well adsorbed to the soil. However, as in the other cases, the extract did make other soil unwettable.

As one would expect, wettable soil can be made unwettable by the application of a number of hydrophobic substances. Thus, if olive oil, palmitic acid, oleic acid, and cholesterol are dissolved in ether and applied to wettable soil in the proportion of one gram to 100 grams of soil, the soil becomes completely unwettable, except in the case of the cholesterol treatment, where the soil can be made to wet by stirring water into the soil. Of interest is that cholesterol is distinguished from the first three substances, which are fatty acids, in that it is a secondary alcohol with several side-chains, including one -OH bond.

An attempt to relate the nature of the particle coatings with the degree of unwettability is shown in Table 1. It will be noted that no clear-cut relationship could be detected.

References

Adams, N. K., Principles of penetration of liquids into solids, *Discussions Faraday Soc.*, **3**, 5–11, 1948.

Bartell, F. E., and H. H. Zuidema, Wetting characteristics of solids of low surface tension such as talcs, waxes and resins, *Am. Chem. Soc.*, **58**, 1449–1454, 1936.

Domingo, W. R., An investigation of irreversibly and difficultly wettable soils (in Dutch), *Landbouwkund. Tijdschr.*, **62**, 252–260, 1950.

van't Woudt, B. D., On factors governing subsurface storm flow in volcanic ash soils, *Trans. Am. Geophys. Union*, **35**, 136–144, 1954.

van't Woudt, B. D., Soil moisture and fertility effects on clover yield, *Soil Sci.*, **80**, 1–9, 1955.

Wander, I. W., An interpretation of the cause of water-repellent sandy soils found in citrus groves in Central Florida, *Science*, **110**, 299–300, 1949.

Zuncker, F., The behavior of water in the soil (in German), *Proc. Second Intern. Congr. Soil Sci.*, Comm. VI, 89–95, 1930.

(Manuscript received October 30, 1958.)

Letter to the Editor

Discussion of Paper by Gast, Kulp, and Long on 'Absolute Age of Early Precambrian Rocks in the Bighorn Basin of Wyoming and Montana, and Southeastern Manitoba'

Frank W. Osterwald

U. S. Geological Survey
Denver, Colorado

Gast, Kulp, and *Long* [1958] presented some valuable data concerning the absolute ages of Precambrian rocks in the Bighorn and Beartooth Mountains of northwestern Wyoming and south-central Montana. In such analytical work, however, geologic evidence, qualitative and inconclusive though it may be, is commonly overlooked or inadequately correlated with the laboratory results. The authors [1958, p. 326] correlated age determinations on greenish-brown biotite separated from a single sample from the Bighorn Mountains with a rather simple sequence of events in the Beartooth Mountains that involves only one interval of regional metamorphism and granitization. The Precambrian history of the Bighorn Mountains probably was much more complex [*Osterwald,* 1955] and the single sample from the Bighorns may reflect any one of four events, or a combination of these events. Furthermore, several lines of evidence indicate that biotite in these Precambrian rocks had a complex history.

The rocks of Precambrian age in the Bighorn Mountains may be generally divided into two series [*Osterwald,* 1955]. Widely different structural patterns of earlier gneisses and schists and later granitic rocks [*Osterwald,* 1958, 1949a, 1949b] indicate that the two rock sequences were formed during two distinct Precambrian deformations. After the second deformation the granitic rocks cooled and became solid enough to fracture along well-defined planes, into which numerous diabase dikes were emplaced [*Darton,* 1906, p. 19–22; *Osterwald,* 1955], some of which were disrupted either by further deep-seated deformation or by remobilization of the granitic rocks [*Osterwald,* 1955, p. 322]. Some rocks are fine grained, highly sheared, silicified, epidotized and chloritized; such rocks occur in elongate zones, and are related to a later episode of Precambrian deformation [*Osterwald,* 1949a].

The sample used for age determinations of Bighorn metamorphism by *Gast, Kulp,* and *Long* [1958, p. 332, Fig. 1] was obtained near the northern edge of the outcrop area of Precambrian rocks. The rock is an augen gneiss, but closely adjacent outcrops in that part of the Precambrian area are extremely variable both in texture and in content of dark minerals. The extreme variation makes very difficult any attempt to correlate the 2760 $\pm$ 70 m y or 2540 $\pm$ 50 m y ages [*Gast, Kulp,* and *Long,* 1958, p. 326] with particular events in the Precambrian history of the Bighorn Mountains. Studies of thin sections cut from rocks similar to that analyzed by them suggest that early biotite and amphiboles were recrystallized at low temperature to epidote, chlorite, hornblende, garnet, and actinolite; this recrystallization implies at least a redistribution of the elements originally contained in the biotite. Furthermore, thin-section studies indicate that biotite is older than potash feldspar porphyroblasts that are of the same age as the granitic rocks. These facts suggest that the present chemical and physical properties of the biotite result from several events in the complex Precambrian history of the Bighorn Mountains. The chemical and physical properties of the greenish-brown biotite analyzed by Gast, Kulp, and Long may have been a palimpsest from the older sequence of gneisses, a result of crystallization or recrystallization at the time of emplacement of the granitic rocks, a result of later downfolding and remobilization of the granitic rocks,

or a result of late Precambrian shearing and recrystallization.

The evidence presented by Gast, Kulp, and Long clearly suggests a widespread, single, major regional metamorphism in Montana and Manitoba, but where this metamorphism fits within the sequence of events in the Bighorns is not clear. Their conclusion [1958, p. 327–328] that ". . . the ages reported here represent a single geologic event, namely, that of *the* (author's italics) major regional metamorphism and granitization . . ." occurred in the Bighorn Mountains as well as in the Beartooth Mountains is not substantiated by the field evidence in the Bighorn Mountains. The conclusion cannot be substantiated because several major geologic events of Precambrian age occurred in the Bighorn Mountains [*Osterwald*, 1949a, 1949b, 1953, 1958], any one of which may represent the single event of the Beartooths.

References

Darton, N. H., Geology of the Bighorn Mountains, *U. S. Geol. Survey Prof. Paper* 51, 1906.

Gast, P. W., J. L. Kulp, and L. E. Long, Absolute age of early Precambrian rocks in the Bighorn Basin of Wyoming and Montana, and southeastern Manitoba, *Trans. Amer. Geophys. Union*, **39**, 322–334, 1958.

Osterwald, F. W., Structure of the Tongue River area, Bighorn Mountains, Wyoming, *Wyo. Geol. Assoc. Guidebook*, 4th Ann. Field Conference, 37–38, 1949a.

Osterwald, F. W., Relation of structure to petrology in the northern Bighorn Mountains, Wyoming (abstract), *Bul. Geol. Soc. Amer.*, **60**, 1965, 1949b.

Osterwald, F. W., Petrology of Precambrian granites in the northern Bighorn Mountains, Wyoming, *J. Geol.*, **63**, 310–327, 1955.

Osterwald, F. W., Structure and petrology in the northern Bighorn Mountains, Wyoming, *Bul. Geol. Survey Wyo.*, **48**, 1958.

(Received July 29, 1958.)

AMERICAN GEOPHYSICAL UNION

1515 Massachusetts Avenue, N.W., Washington 5, D. C.

Established by the National Research Council in 1919 for the development of the science of geophysics through scientific publication and the advancement of professional ideals.

APPLICATION FOR MEMBERSHIP

Please refer to qualifications on reverse side and designate below type of membership desired:

Member ($10) ☐ Associate ($10) ☐ Student ($3) ☐

Application forms for Corporation Membership are available upon request.

1. _____
Surname *First Name* *Middle Name*

2. _____
Preferred mailing address for publications

_____
Permanent address

3. _____ 4. _____
Place *Month* *Day* *Year of Birth* *Country of citizenship/naturalization*

5. _____
Nature of work and title and/or military rank; name and address of organization with which you are associated.

6. Check section or sections with which affiliation is desired.

☐ Geodesy
☐ Seismology
☐ Meteorology
☐ Geomagnetism and Aeronomy
☐ Oceanography
☐ Volcanology, Geochemistry, and Petrology
☐ Hydrology
☐ Tectonophysics

7. **EXPERIENCE (List below)**

Dates: From To Name and address of organization Title, duties, nature of work

8. **EDUCATION (List below)**

Dates: From To School Address Major Subject Degree, if any

*9. References: Please list below names and addresses of two or three references; include members of the AGU or others who know you well.

*10. Titles of technical contributions or publications, particularly those in the geophysical sciences, and where published.

*11. Brief statement of any special interests or qualifications in the geophysical sciences.

Date_____ _____
Written Signature

* Applicants for student membership may omit Questions 9, 10, and 11, but must fill in Question 12. Please return form with check or money order payable to American Geophysical Union, 1515 Massachusetts Ave., N.W., Washington 5, D. C.

(over)

12. (STUDENT MEMBERS ONLY) The person whose signature appears on the reverse side is known to me and is a student majoring in_____(subject) at_____ (Name of college or university) expected to graduate in_____(year) with the degree of_____

☐ He is a full-time student, or ☐ a teaching or research assistant enrolled in more than half of a full-time academic program.

_____
(Signature of faculty sponsor)

☐ Check here if faculty sponsor is a member of AGU and willing to act as a regular sponsor for associate membership as well.

_____
(Typed or printed name of sponsor)

_____
(Title)

QUALIFICATIONS FOR MEMBERSHIP IN THE AMERICAN GEOPHYSICAL UNION

The membership of the AGU shall consist of Members, Associate Members, Student Members, and Corporation Members.

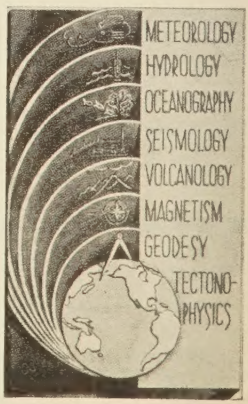
Those eligible as candidates for election to the grade of MEMBER shall be:

MEMBER (a) Persons who have made an active contribution to geophysical research through observation, publication, teaching, or administration. Definite evidence should be presented to the Membership Committee. "Publications" may include books, articles, unpublished manuscripts, inventions, or development of geophysical instruments.

(b) Persons who have made active practical application of geophysical research. It should be shown that the nominee's work has not been purely routine, but that it has tended to create new knowledge of, or to broaden or strengthen the application of, geophysical research. In general, the minimum qualifications for membership will be not less than three years of professional experience in some phase of geophysics.

Those eligible as candidates for election to the grade of ASSOCIATE MEMBER shall be:

ASSOCIATE MEMBER Persons who have an active interest in physical processes of the Earth or technical assistance in the application of geophysics. In general, the minimum qualification for associate membership will be acceptable training or experience in some field of geophysics or allied science.



CORPORATION MEMBER Corporations and other interested organizations shall be eligible as candidates for election to CORPORATION MEMBERSHIP. They shall have the privilege of designating a representative who has the rights and privileges of Members (use special form).

STUDENT MEMBER Those eligible as candidates for election to the grade of STUDENT MEMBER shall be persons who are graduate or undergraduate students in residence at least half-time and who are specializing in the geophysical sciences. Teaching or research assistants enrolled in more than half of a full-time academic program may also be eligible for Student Membership. Student Members shall have all the privileges of Members except that they shall not vote or hold office.

(Journal of Geophysical Research, Vol. 64, No. 2, Feb., 1959)

Contents